PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/12, C07K 14/47, C12Q 1/68, A01K 67/027, A61K 38/17, G01N 33/68, C07K 16/18, C12N 5/10, 15/62, A61K 48/00

A2

(11) International Publication Number: WO 97/01634

(43) International Publication Date: 16 January 1997 (16.01.97)

(21) International Application Number: PCT/IT96/00131

(22) International Filing Date: 27 June 1996 (27.06.96)

(30) Priority Data:
RM95A000434 27 June 1995 (27.06.95) IT

(71) Applicant *(for all designated States except US)*: ISTITUTO DI RICERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S.P.A. [IT/IT]; Via Pontina Km 30.600, I-00040 Pomezia (IT).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: JIRICNY, Josef [US/IT]; Via Talete, 41, I-00124 Casalpalocco (IT). PALOMBO, Fabio [IT/IT]; Via Gran Sasso, 10/A, I-00141 Rome (IT). GALLINARI, Paola [IT/IT]; Piazza S. Maria Liberatrice, 18, I-00153 Rome (IT).

(74) Agents: DI CERBO, Mario et al.; Società Italiana Brevetti S.p.A., Piazza di Pietra, 39, I-00186 Roma (IT).

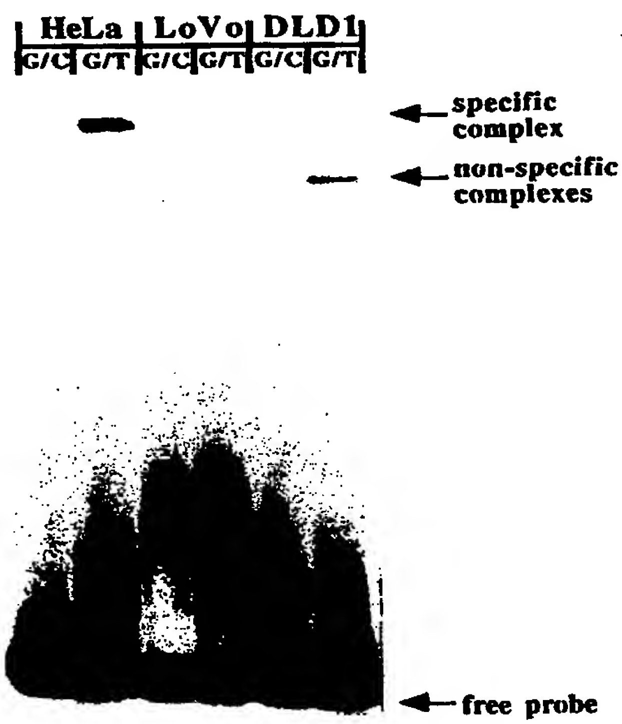
(81) Designated States: AU, CA, CN, JP, MX, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: POLYPEPTIDE FOR REPAIRING GENETIC INFORMATION, NUCLEOTIDIC SEQUENCE WHICH CODES FOR IT AND PROCESS FOR THE PREPARATION THEREOF (GUANINE THYMINE BINDING PROTEIN - GTBP)

(57) Abstract

The present invention relates to a new protein, GTBP (Guanine Thymine Binding Protein), that binds to G/T DNA mismatches to mediate repair of genetic information, to methods for detection of this protein, to the nucleotidic sequence encoding this protein and to processes for obtaining the above-mentioned protein using genetic engineering techniques. Furthermore, the present invention has as its object the detection in tumor tissues of the mutant GTBP gene in order to prevent and provide rapid diagnosis of human colorectal tumor forms. The figure shows the absence of GTBP-specific activity in cells obtained from human colorectal tumors.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

POLYPEPTIDE FOR REPAIRING GENETIC INFORMATION, NUCLEOTIDIC SEQUENCE WHICH CODES FOR IT AND PROCESS FOR THE PREPARATION THEREOF (GUANINE THYMINE BINDING PROTEIN - GTBP).

DESCRIPTION

Technical field

This invention relates to the area of cancer prevention, diagnosis and therapeutics. In particular, the invention is concerned with methods for detection of a novel mismatch binding protein, termed GTBP (Guanine Timine Binding Protein), which mediates the repair of genetic information, with the nucleic acid sequence encoding the protein and with processes for obtaining the protein and producing it by recombinant genetic engineering techniques. In addition, the present invention also relates to detection of mutated GTBP gene in tumour tissues and to prevention and early diagnosis of human colorectal cancers.

Background of the discovery

In human cells, mismatch recognition and binding has until now been believed to be mediated by the hMSH2 protein. The observation that cells from human colorectal cancers (CRC) exhibit a mutator phenotype with a marked instability of microsatellite sequences suggested that these tumor cells may be deficient in DNA mismatch repair. This hypothesis was substantiated when extracts from CRC tumor-derived cell lines where shown to be unable to repair mismatches in an *in vitro* assay (see refs. 1 and 2 for reviews).

The serendipitous discovery of an open reading frame (ORF) encoding a polypeptide homolog of the E. *coli* mismatch-binding protein MutS (3, 4) paved the way for the identification of an ever-growing family of MSH genes, ranging from bacteria to man (see e.g. 5). Three members of this family, *S. cerevisiae* MutS homologs MSH1 and MSH2, as well as the human homolog hMSH2, could be shown to bind to mismatched DNA *in vitro* (6-9). The link between the biological function of hMSH2 and the

phenotype of the CRC tumors was forged when (i) the *hMSH2* gene was shown to segregate with a known CRC locus on chromosome 2p (10,11), (ii) the *hMSH2*-deficient cell line LoVo was shown to be deficient in mismatch repair (12) as well as in mismatch-binding activity (13) and (iii) the genome of this cell line exhibited a marked instability of microsatollite sequences (14). A mismatch-binding factor, GTBP (for G/T binding protein), originally identified in HeLa cells by the present inventors (15), was shown to bind preferentially to heteroduplexes containing G/T mispairs. Purification of this DNA binding activity by G/T mismatch affinity chromatography yielded a mixture of two polypeptides of apparent molecular weights of 100 and 160 kDa (16), indicating that the mismatch-specific complex was composed of two proteins. The 100 kDa constituent of the complex was demonstrated to be hMSH2 (17). The present discovery implies that hMSH2 acts as a complex with GTBP in the correction of base/base mispairs and one- or two-nucleotide loops. Moreover, GTBP is necessary but not indispensible in the correction of larger insertion/deletion loops. A number of tumors have been shown to display mutator phenotypes which are consistent with the functional role of the hMSH2-GTBP complex (20-24). Prior to the current discovery and characterization of GTBP, no specific role in the repair of genetic information and no hereditary defect had been associated with this protein or with the gene encoding it.

Relevant Literature

1. P. Modrich, *Science* **266**, 1959 (1994)

2. J. Jiricny, *Trends Genet*. **10**, 164, 1994.

3. J. P. Linton et al. *Mol.Cell. Biol* 9, 303(1989)

4. H. Fujii and T. Shimada, *J. Biol.Chem.* **264**, 10057 (1989).

5. L. New, K.Liu, G.E. Crouse, *Mol. Gen. Genet.* **239**, 97 (1993).

6. N.-W. Chi and R. D. Kolodner, *J. Biol. Chem.* **269**, 29984 (1994).

7. T. Prolla *et al, Science* **265**,1091(1994).

8. B. Alani, N-W Chi, R. D. Kolodner, *Genes & Development* 9, 234 (1995).

9. R. Fishel, A. Ewel, M.K. Lescoe, *Cancer Research* **54**, 5539 (1994).

10. R. Fishel *et al., Cell* 75,1027 (1993).

11. F.S. Leach *et al., Cell* 75,1215 (1993).

12. A. Umar *et al., J. Biol. Chem.* **269**, 14367 (1994).

13. G. Aquilina *et al., Proc. Natl. Acad. Sci. U.S.A.*. **91**, 8905 (1994).

14. D. Shibataet *al, Nature Genet.* **6**, 273 (1994).

15. J. Jiricny *et al. Proc. Natl. Acad. Sci. U.S.A.* **85**, 8860 (1988).

16. M. Hughes and J. Jiricny, *J. Biol. Chem.* **267**, 23876 (1992).

17. F. Palombo *et al. Nature* **367**, 417 (1994).

18. J. Lingner, J. Kellernan and W. Keller, *Nature* **354**, 496 (1991).

19. Da. Costa *et at., Nature Genetics* **9**, 10 (1995).

20. L.A. Aaltonen *et al., Science* **260**, 812 (1993)

21. S.N. Thibodeau *et al., Science* **260**, 816 (1993)

22. Ionov *et al., Nature* **363**, 558 (1993)

23. R. Wooster *et al., Nature Genetics* 6, 152 (1994)

24. A. Merlo *et al., Cancer Res.* **54**, 2098 (1994).

25. J.D. Dignam *et al., Methods Enzymol.* **101**, 382 (1983).

26. P. Gallinari *et al., J. Virol.* **68**, 3809 (1994).

Sumary of the invention

It is an object of the present invention to provide a 1360-amino acid sequence corresponding to the polypeptide referred to as GTBP. It should be stated that GTBP is used to indicate a compound polypeptide combining in order the amino acid sequences indicated in SEQ ID NO:15 (from amino acid 1 to 68) and SEQ ID NO:1 (from amino acid 1 to 1292).

It is another object of the present invention to provide a genetic construct containing a double-stranded cDNA sequence of 4080 base pairs encoding a 1360-amino acid peptide referred to as GTBP. It should be stated that the whole coding gene GTBP indicates a compound DNA sequence combining in order the nucleotide sequences indicated in SEQ ID NO:16 (from nucleotide 1 to 204) and SEQ ID NO:12 (from nucleotide 1 to 3980).

A further object of the present invention is to provide a genetic construct capable of expressing a 1360-amino acid peptide of molecular mass 153 kDa referred to as GTBP.

It is another object of the present invention to provide a method for preparation and isolation of native GTBP protein in pure form from cultured cells and tissues.

It is another object of the present invention to provide a method for the assessment of the *in vitro* activity of GTBP.

It is yet another object of the present invention to provide a method for the detection of mutated GTBP by the use of specific antibodies directed against GTBP.

It is yet another object of the present invention to provide a method for the detection of mutated *GTBP* alleles by the use of the polymerase chain reaction and sequencing of the amplification products.

It is another object of the present invention to provide DNA probes for the detection of mutated *GTBP* genes in human cells.

It is an object of the present invention to provide a method for diagnosing and prognosing of human colorectal cancers (CRC).

It is yet another object of the present invention to provide a method for detecting the genetic predisposition to human colorectal cancers (CRC).

It is yet another object of the present invention to provide a method for large-scale population screening to genetic predisposition to human colorectal cancers (CRC).

It is still another object of the present invention to provide a method for supplying wild-type *GTBP* alleles to a cell which has lost the *GTBP* gene function.

It is another object of the present invention to provide a method for generating transgenic animals carrying mutant *GTBP* alleles.

It is another object of the present invention to provide a method for testing the activity of therapeutic agents aimed to suppress human colorectal cancers (CRC).

These and other objects of the invention are provided by one or more of the embodiements which are described below.

In one embodiment the sequence of a 1360-amino acid polypeptide is provided corresponding to the protein referred to as GTBP.

In another embodiment a cDNA molecule is provided which comprises the coding sequence of the *GTBP* gene.

In another embodiment a procedure for the preparation of the pure GTBP protein is provided.

It is another embodiment of the present invention to provide pairs of single stranded primers to determine the nucleotide sequence of the *GTBP* gene or of DNA regions internal to the *GTBP* gene by polymerase chain reaction. The sequence of said primers is internal to chromosome 2p16, said pairs of primers allowing the syntesis of *GTBP* gene or of parts of it.

In yet another embodiment of the present invention a nucleic acid probe is provided which is complementary to human wild-type *GTBP* gene coding sequence and which can form mismatches when annealed with mutant *GTBP* alleles, thereby making possible the detection of heteroduplex DNA as revealed by shifts in electrophoretic mobility either with or without prior enzymatic or chemical cleavage.

In another embodiment a procedure is indicated for the detection of wild-type or mutated GTBP protein in humans, comprising: isolating a human sample selected from the tissue or body fluid and detecting the wild-type or the altered GTBP protein itself or in any complex formed by the association of GTBP with other polypeptides.

In another embodiment of the present invention a method is provided for the assessment of the activitiy of (i) the wild-type GTBP protein or (ii) of derived peptides obtained by deletion or insertion of known amino acid sequences in GTBP protein or (iii) of the altered GTBP protein as the result of *in vivo* mutational events or (iv) of any complex formed by the association of peptides just mentioned in (i), (ii), (iii), and (iv) of the present embodiment with other polypeptides.

In yet another embodiment a method is provided for the detection of cancer in humans, comprising: isolating a human sample selected from the tissue or body fluid; detecting the alteration in the *GTBP* gene or in the expressed polypeptide (GTBP protein) itself or in any complex formed by the association of GTBP with other polypeptides, said alteration indicating the predisposition to neoplastic transformation or the presence of cancer.

In still another embodiment of the present invention a method of diagnosing or prognosing neoplastic tissue of a human is provided comprising: detecting somatic alterations in wild-type *GTBP* alleles or their expression products in human colorectal cancers (CRC), said alteration indicating neoplasia of the tissue.

In yet another embodiment a method is provided for the detection of genetic predisposition to CRC, comprising: isolating a human sample selected from the group consisting of blood, bioptic samples of tissues, esfoliative cells and any other generic human sample; detecting the alteration in the *GTBP* gene or in the

expressed polypeptide (GTBP protein) itself or in any complex formed by the association of GTBP with other polypeptides, said alteration indicating genetic predisposition to cancer.

In another embodiment of the present invention a method is provided for supplying wild-type *GTBP* gene function to a cell which has lost said gene function by virtue of any mutation in the *GTBP* gene, comprising: introducing wild type *GTBP* gene into a cell which has lost said gene function such that *GTBP* gene is then expressed at wild-type level in the cell. GTBP protein can also be applied to cells or administered to animals to remediate defects in *GTBP* gene function.

In an additional embodiment a method is provided to supply a portion of wild-type *GTBP* gene to a cell which has lost the said gene such that the said portion is expressed in the cells and encodes part of the GTBP protein which is required for non-neoplastic growth of the said cell.

It is another embodiment of the present invention the generation of transgenic animals carrying a mutated *GTBP* gene derived from a second species or a mutated *GTBP* gene generated *in vitro* by genetic engineering techniques.

In another embodiment of the present invention a method of testing therapeutic agents for the ability to suppress a neoplastically trasformed phenotype is provided. The method comprises: applying a test substance to a cultured epithelial cell which carries a mutation of the *GTBP* gene and determining whether the substance suppresses the neoplastic phenotype of the cell or suppresses the growth of already developed tumors.

In another embodiment of the present invention a method of testing therapeutic agents for the ability to suppress a neoplastically trasformed phenotype is provided. The method comprises: applying a test substance to an animal which carries a mutation of the *GTBP* gene

and determining whether the substance prevents neoplastic transformation of defined tissues or suppresses the growth of already developed tumors.

The present information provides the art with the information that the *GTBP* gene, a heretofore unknown gene, encodes the GTBP protein which acts as specific mismatch-binding factor. GTBP binds preferentially to heteroduplexes containing G/T mispairs and one- or two-nucleotide loops. Purification of this DNA binding activity made it possible to establish that the mismatch-specific factor is in fact a complex composed of two distinct proteins. The smaller constituent of the complex (about 100 kDa) is the hMSH2 protein (17) whereas the larger component (about 160 kDa) is GTBP. The present invention provides the technical tools for the detection and for the activity assessment of GTBP alone or as a complex with hMSH2. The *GTBP* gene is a target of mutational events, these alterations being associated with tumorigenesis. This discovery allows highly specific assays to be performed to determine the neoplastic status of a particular tissue or the predisposition to cancer of individuals. A number of tumors have been shown to display mutator phenotypes with a similarly low degree of microsatellite instability (20-24) consistent with the functional role of the hMSH2-GTBP complex. Prior to the current discovery and characterization of GTBP, no specific role in the repair of genetic information and no hereditary defect had been associated with this protein.

Brief description of the drawings.

Figure 1 a shows the commercial phagemid vector pBluescript SK$^-$ (Stratagene) used for cloning and sequencing the GTBP cDNA. The DNA fragment shown in SEQ ID NO: 12 was cloned between the *Eco*RI and *Xho*I sites of the vector. **b** shows the commercial pCITE 2b vector. The insert described in SEQ ID NO: 12 was inserted between the *Eco*RI and *Xho*I sites of the vector.

Ampicillin = beta-lactamase gene for ampicillin resistance
ColE1 ori = origin of replication derived from plasmid ColE1
f1 = origin of replication of phage F1
lacZ = alpha peptide of beta-galactosidase used for genetic complementation
MCS = multiple cloning site containing the recognition sequences of the listed restriction enzymes
T3 and T7 = promoter sequences from phages T3 and T7.

Figure 2 shows the commercial plasmid vector pGEX-3x (Pharmacia Biotech) that was used for cloning of the PCR fragments corresponding to amino acid residues 27 to 158 of hMSH2 and 750 to 928 of GTBP (SEQ ID NO:1). Primers used for amplification were: 5'CGGGATCCCCCCGGAGAAGCCGACCACCAC3' and 5'CGGAATTCCTGGCCATCAACTGCGGACAT3' for codons 27 to 158 of hMSH2, and 5'CGGAATTCTCAACTCGTATTCTTCTG3' and 5'CGGGATCCCCCTTGAGAGGCTACTCAGT3' for codons 750 to 928 of GTBP. The PCR products, identified respectively as SEQ ID NO: 13 and 14 were cloned between the *Bam*HI and *Eco*RI sites. The expression products, in the form of polypeptides fused with glutathione-S-transferase, were purified by affinity chromatography on a commercial glutathione matrix (Pharmacia Biotech) as directed by the manufacturer. The pure fusion proteins were used for the immunization of New Zealand White SPF female rabbits by standard protocols as reported in the publication *Antibodies: A Laboratory Manual* (1988) eds. Harlow and Lane, Cold Spring Harbor Laboratories Press.

Figure 3 shows an alignment of the amino acid sequences of the conserved C-terminal regions of the four mismatch binding proteins, i.e. GTBP (*H. sapiens*), hMSH2 (*H. sapiens*), MSH2 (*S. cerevisiae*) and MutS (*E. coli*). Identical residues are in black boxes, conserved ones in shaded boxes. Sequences reported in the alignment correspond to entries MSH2_YEAST (MSH2) and MUTS_ECOLI

(MutS) in the SwissProt databank, or the coding region of GenBank entry HSU04045 (hMSH2). The alignements show that a high degree of conservation exists among the three homologs, with the C-terminal part of the protein being particularly highly conserved. GTBP can therefore be considered a new member of the MSH family.

Figure 4 shows the sequence homology, at the protein level, between pairs of MSH family members. **Section a** shows the matrix obtained from the alignment of GTBP (on the abscissa) with the yeast GTBP homolog (GenBank accession number Z47746, on the ordinate); the two proteins show comparable length and a significant homology is evident throughout their whole sequence. **Section b** shows the matrix obtained from the alignment of yeast MSH2 (on the ordinate) with GTBP (on the abscissa); the proteins show different lengths and most of the homology is confined to the C-rerminal regions of the two sequences. **Section c** shows the matrix obtained from the alignment of human MSH2 protein (on the ordinate) with GTBP (on the abscissa); the proteins show different lengths and, also in this case, most of the homology is confined to the C-rerminal regions of the two sequences. **Section d** shows the matrix obtained from the alignment of human hMSH2 protein (on the ordinate) with the yeast MSH2 (on the abscissa); the two proteins show comparable length and the homology is evident throughout the entire sequence.

Figure 5 shows the effect of selective anti-hMSH2 and anti-GTBP antisera on the formation of the specific mismatch-binding complex. Pre-incubation of HeLa nuclear extracts with either antiserum prior to addition of the G/T heteroduplex DNA probe results in a diminuition of the specific band in the gel-shift assay, an effect not observed when the respective pre-immune sera were used. This figure proves that both hMSH2 and GTBP are present in the mismatch-binding factor. This gel-shift analysis was carried out as described in ref.15, except that

nuclear extracts were used (25). The antisera were added to the reaction mixtures 20 min prior to the addition of the radioactively-labelled probe. The figure is an autoradiogram of a native 6% polyacrylamide gel run in Tris-acetate/EDTA (TAE) buffer prepared according to Maniatis *et al.*, *Molecular cloning: a laboratory manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982.

Figure 6 shows that the mismatch-binding activity can be reconstituted using GTBP and hMSH2 obtained using an *in vitro* translation system. The procedure followed to generate *in vitro* transcripts of the hMSH2, C1 and FLY5 coding sequences was as follows: The DNA region encoding hMSH2 was inserted into pCite-1; C1 and FLY5 ORFs were introduced into pCite-2b (Novagen). *In vitro* transcription and translation reactions were carried out as described in ref. 26, including a mock translation reaction in the absence of added DNA. ^{35}S-labeled translation products were analysed on a SDS-polyacrylamide gel treated with Amplify (Amersham), dried and autoradiographed. Gel-shift assays were performed as described in ref. 15. Aliquots of 5 µl of the single *in vitro* translation reactions were tested; in the pre-mixing experiments, 2.5 µl of each of the two translation reactions were mixed and incubated for 15 min at room temperature before the addition of the probe. AMP at a concentration of 5 mM was included in all the DNA binding reactions so as to overcome the effect of ATP in the reticulocyte lysates, which prevents the formation of mismatch-specific protein-DNA complexes, according to ref. 16. **Section a** is an autoradiogram of a denaturing 7.5% SDS-polyacrylamide gel showing that translation of hMSH2, GTBP (C1) and FLY5 mRNAs in a reticulocyte lysate system (Promega) gave rise to expected polypeptides of 113, 142 and 122 kDa, respectively. **Section b** shows the gel-shift analysis which demonstrates the binding of the *in vitro*-translated proteins to the G/T heteroduplex. The

figure is an autoradiogram of a native 6% polyacrylamide gel run in TAE buffer.

Figure 7 shows that mismatch binding activity is absent from cell extracts lacking GTBP or hMSH2. The experiment is based on the analysis of two cell lines derived from CRC: LoVo cells contain a homozygous deletion of *hMSH2* alleles and do not exhibit G/T binding activity (13), while neither *hMSH2* allele is mutated in DLD1 cells, in spite of the fact that also this cell line lacks G/T binding activity. **Section a** shows a gel-shift assay showing that extracts of LoVo and DLD1 fail to make mismatch-specific complexes. The G/C and G/T probes were obtained as described previously (15). Experimental conditions were as in Figure 6. The figure is an autoradiogram of a native 6% polyacrylamide gel run in TAE buffer. **Section b** shows the Western blot analysis of extracts from Hela, LoVo and DLD1 cells. The protein bands were visualized using an alkaline phosphatase-conjugated anti-rabbit IgG system (Promega) as directed by the manufacturer. In the two left lanes, the anti-GTBP and anti-hMSH2 antisera were used alone with the HeLa extract to demonstrate their selectivity for the 160 and 100 kDa proteins, respectively. In the remaining lanes, both antisera were used together. Control HeLa cells revealed the presence of both hMSH2 and GTBP. In contrast, the two CRC-derived tumor cell lines LoVo and DLD1 were completely devoid of full-length hMSH2 and GTBP, respectively. The amounts of hMSH2 in DLD1 cells and GTBP in LoVo cells were considerably lower than in HeLa cells. Since hMSH2 and GTBP bind heteroduplex DNA as a complex, the lack of one of the two proteins may cause instability of the second component of the complex.

Figure 8, part a, shows the experimental approach followed to discover the amino-terminal region of GTBP (from amino acid 1 to 68 of SEQ ID NO:15). Using the 5' RACE method(Rapid Amplification cDNA Ends, given in detail in the publication Nicolaides, N.C. et al.

Genomics, 29: 229-234, 1995 and Nicolaides N.C. et al. *Genomics*, 30: 195-206, 1995) it is possible to determine the sequence upstream of the amino acid Ala in position 1 of SEQ ID NO:1. Initially, a pair of oligonucleotides was used that pairs with the sequence given in SEQ ID NO:12 from nucleotide 114 to 133 (primary oligonucleotide A) and from nucleotide 56 to 74 (secondary oligonucleotide B). The PCR reaction products were sequenced and it was possible to determine that the amplification product was capable of encoding the polypeptide DAAWSEAGPGPR, corresponding to amino acids 46-58 of the amino-terminal domain of GTBP as indicated in SEQ ID NO:15. Using a further two oligonucleotides, whose sequence was deduced from the initial RACE, complementary to the sequence given in SEQ ID NO:16 from nucleotide 188 to 204 (primary oligonucleotide C) and from oligonucleotide 169 to 185 (secondary oligonucleotide D) it was possible to amplify the GTBP-coding region 5' by-passing the methionine in position 1 of the amino acid sequence given in SEQ ID NO:15. The amplified clone, termed KMN, contained the entire nucleotidic sequence given in SEQ ID NO:16. RACE analysis of leucocyte cDNA is shown in lanes 2 and 5, that of placenta cDNA in lanes 3 and 6. The products of lanes 1 to 3 derive from sequenced amplifications with oligonucleotides A and B, those in lanes 4 to 6 derive from sequenced amplifications with oligonucleotides C and D. Lanes 1 and 4 are the negative controls (absence of template). The molecular weight markers are indicated at the side.

Part b of figure 8 shows expression of the transcript encoding the protein GTBP using RT-PCR (PCR preceded by inverse transcription on RNA templates). The RT-PCR was carried out using a synthetic oligonucleotide which paired with the sequence given in SEQ ID NO:12 from nucleotide 114 to 133 in the inverse transcription reaction followed by amplification with an

oligonucleotide with a sequence equal to the end 5' of the GTBP transcript, that is 5'GGTGCTTTTAGGAGCCCCG3'.

The RNA used as a mold template taken from HeLa cells (lane 2) placenta (lane 3) leucocytes (lane 4) and cells from the colon (lane 5); these were incubated with (+ symbol on the lane) or without (- symbol on the lane) inverse transcriptase and then made to underto PCR. Wnere no cDNA was produced, as the reverse transcription reaction did not occur, it was not seen to be amplified. Lane 1 is the negative control without RNA.

Detailed description

In view of the potential and varied roles for mismatch binding proteins in the repair of genetic information and their effects on disease state, such as tumor cell transformation and proliferation, metastases, and the paucity of understanding of the molecules and agents that selectively effect or modulate the activities of these proteins there exists a need in the art for compounds and agents with effector and modulator activity and methods to identify these and related compositions and agents. Further, such agents can serve as commercial research reagents for control of nucleic acid repair, and other GTBP-related conditions. Despite progress in developing a more defined model of the molecular mechanisms underlying nucleic acid repair, few significant methods applicable to assessing predisposition to cancer and or to its treatment have evolved. The hMSH2/GTBP heterodimer is necessary for the correction of base/base mispairs and one or two-nucleotide loops. Genomic instability in tumor-derived cell-lines lacking GTBP demonstrates itself mainly in the form of small differences (e.g. in runs of A) rather than large changes in CA repeats, characteristic of phenotypes associated with the four known CRC loci *hMSH2*, *hMLH1*, *hPMS1* and *hPMS2*. Cancers displaying mutator phenotypes with a low degree of microsatellite instability (20-24) may be associated with a malfunction of GTBP. It is a

discovery of the present invention that mutational events associated with tumorigenesis in CRC are due to defects in the *GTBP* gene.

Novel compositions comprising generic sequences encoding the GTBP protein, as well as fragments derived therefrom are provided, together with recombinant proteins produced using the genomic sequences and methods of using these compositions.

Exemplary amino acid and DNA sequences of the invention are set forth in SEQ ID NO: 1 - SEQ ID NO:15 and in SEQ ID NO: 12 - SEQ ID NO: 16. Standard abbreviations for nucleotides and amino acids are used in the Figures and elsewhere in this specification. GTBP-derived polypeptides are particularly preferred embodiments of the invention, although variations based on the specific sequences of these polypeptides are also part of the present invention. In its broader aspects, the invention (as it pertains to polypeptides *per se*) includes any polypeptide selected from the group consisting of:

(i) any protein having an amino acid sequence which is at least 85% homologous to the amino acid sequences of SEQ ID NO: 1, SEQ ID NO:15 and the combination thereof, and,

(ii) fragments thereof comprising at least 10 consecutive amino acids located within the amino acid sequences of SEQ ID NO: 1, SEQ ID NO:15 and the combination thereof, wherein the polypeptide is capable of binding to an antibody specific for GTBP.

In the genetic engineering aspects of the present invention, specific coding sequences as set forth in SEQ ID NO: 12, SEQ ID NO:16 and the combination thereof, which correspond to the preferred polypetides are themselves preferred.

Equivalent and complementary DNA and RNA sequences (see below for definitions of these terms) are likewise preferred. In its broader aspects, the genetic engineering aspects of the present invention include any

recombinant DNA or RNA molecule comprising a DNA sequence encoding GTBP itself or GTBP-derived protein according to SEQ ID NO: 1 or a corresponding DNA or RNA sequence, or a subsequence thereof comprising at least 10 nucleotides. The present invention also focuses on diagnostic methodologies aimed to detect loss of GTBP function in humans and consequent predisposition to neoplasia.

Defintion of terms

A number of terms used in the art of genetic engineering and protein chemistry are used herein with the following defined meanings.

Two nucleic acid fragments are "homologous" if they are capable of hybridizing to one another undcr hybridization conditions described in Maniatis *et al.*, (1982), *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., pp. 320-325. By using the following wash conditions --2 x SSC, 0.1% SDS, room temperature twice, 30 minutes each; then 2 x SSC, 0.1% SDS, 50° C once, 30 minutes; then 2 x SSC, room temperature twice, 10 minutes each-- homologous sequences can be identified that contain at most about 25-30% base pair mismatches. More preferably, homologous nucleic acid strand contains 15-25% base pair mismatches, even more preferably 5-15% base pair mismatches. These degrees of homology can be selected by using more stringent wash conditions for identification of clones from gene libraries (or other sources of genetic material), as is well known in the art.

Two amino acid sequences are homologous if there is a partial or complete identity between their sequences. For example, 85% homology means that 85% of the amino acids are identical when the two sequences are aligned for maximum matching. Gaps (in either of the two sequences being matched) are allowed in maximizing matching gap lengths of 5 or less are preferred with 2 or less being more preferred.

Alternatively and preferably, two protein sequences (or polypeptide sequences derived from them of at least 30 amino acids in length) are homologous, as this term is used herein, if they have an alignment score of more than 5 (in standard deviation units) using the program ALIGN with the mutation data matrix and a gap penalty of 6 or greater (Dayhoff, M.O., in *Atlas of Protein Sequence and Structure*, 1972, volume 5, National Biomedical Research Foundation, pp. 101-110, and Supplement 2 to this volume, pp. 1-10). The two sequences or parts thereof are more preferably homologous if their amino acids are greater than or equal to 50% identical when optimally aligned using the ALIGN program.

A DNA fragment is "derived from" a GTBP-encoding DNA sequence if it has the same or substantially the same base pair sequence as a region of the coding sequence for GTBP protein molecule.

"Substantially the same" means, when referring to biological activities, that the activities are of the same type although they may differ in degree. When referring to amino acid sequences, "substantially the same" means that the molecules in question have similar biological properties and preferably have at least 85 % homology in amino acid sequences. More preferably, the amino acid sequences are at least 90% identical. In other uses, "substantially the same" has its ordinary English language meaning.

A protein is "derived from" GTBP if it has the same or substantially the same amino acid sequence as a region of the GTBP protein molecule. By polypeptide derivatives of GTBP protein is meant polypeptides differing in length from the natural protein and containing five or more amino acids in the same primary order as found in the protein as obtained from a natural source. Polypetide molecules having substantially the same amino acid sequence as the natural protein but possessing minor amino acid substitutions which do not significantly

affect the ability of the protein or polypeptide to interact with protein-specific molecules, such as antibodies and nucleic acids are within the definition as derived from GTBP. Derivatives include glycosylated forms, aggregative conjugates with other protein molecules and covalent conjugates with unrelated chemical moieties. Covalent derivatives are prepared by linkage of functionalities to groups which are found in the amino acid chain or at the N-or C-terminal residue by means known in the art.

GTBP-specific molecules include polypeptides such as antibodies that are specific for the protein or polypeptide containing the naturally occurring GTBP amino acid sequence. By "specific binding polypetide" are intended polypeptides that bind with GTBP protein and its derivatives and which have a measurably higher binding affinity for the target polypeptide than for other polypetides tested for binding. Higher affinity by a factor 10 is preferred, more preferably by a factor of 100. Binding affinity for antibodies refers to a single binding event (i.e., monovalent binding of an antibody molecule). Specific binding by antibodies also means that binding takes place at the normal binding site of the molecule's antibody (at the end of the arms in the variable region).

As discussed above, minor amino acid variations from the natural amino acid sequence of GTBP protein are contemplated; in particular, conservative amino acid replacements are contemplated. Conservative replacements are those that take place within a family of amino acids that are related in their side chains. Genetically encoded amino acids are generally divided into four families: (I) acidic = aspartate, glutamate; (2) basic = lysine, arginine, histidine; (3) non-polar = alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan; and (4) uncharged polar, = glycine, asparagine, glutamine, cystine, serine,

threonine, tyrosine. Phenylanine, tryptophan, and tyrosine are sometimes classified jointly as aromatic amino acids. For example, it is reasonable to expect that an isolated replacement of a leucine with an isoleucine or valine, an aspartate with a glutamate, a theonine with a serine, or a similar replacement of an amino acid with a structurally related amino acid will not have a major effect on the binding properties of the resulting molecule, especially if the replacement does not involve an amino acid at a binding site involved in the interaction of GTBP or its derivatives with an antibody or with a specific DNA recognition sequence. Whether an amino acid change results in a functional peptide can readily be determined by assaying the specific binding properties of the polypeptide derivative.

Isolation of cDNA encoding GTBP protein

Isolation of nucleotide sequences encoding GTBP protein involves creation of a cDNA library prepared from full-length mature messenger RNA extracted from cultured cells or tissues. Evidence is provided that GTBP is conserved over a broad evolutionary range, thus allowing the isolation of GTBP homologs from the genomes of phylogenetically distant species, i.e. from mammals to yeasts to bacteria.

Genetic libraries can be made in either eukaryotic or prokaryotic host cells. Widely available cloning vectors such as plasmids, cosmids, phage, YACs and the like can be used to generate genomic libraries suitable for the isolation of nucleotide sequences encoding GTBP protein or portions thereof. Useful methods for screening genetic libraries for the presence of GTBP protein nucleotide sequences include the preparation of oligonucleotide probes based on the sequence information provided in SEQ ID NO: 1 and SEQ ID NO: 15 (after decoding of the amino acid sequence) as well as in SEQ ID NO:12 and SEQ ID NO: 16 (directly derived from the encoding DNA) of this patent. By employing the standard

triplet genetic code, oligonucleotide sequences of about 17 base pairs or longer can be prepared by conventional *in vitro* synthesis techniques. The resultant nucleic acid sequences can be subsequently labeled with radionuclides, enzymes, biotin, fluorescers or the like, and used as probes for screening the libraries.

Additional methods of interest for isolating GTBP protein-encoding nucleic acid sequences include screening of genetic libraries for the expression of GTBP protein or fragments thereof by means of GTBP protein-specific antibodies, either polyclonal or monoclonal. Moreover, a selection method advisable for the screening of GTBP libraries cloned in conventional expression vectors is based on the specific binding of the protein (or of polypeptides contained therein) to heteroduplex DNA molecules containing G/T mimatches. A particularly preferred technique for isolating homolog proteins from related species or strains involves the use of degenerate primers based on partial amino acid sequences of GTBP protein and the polymerase chain reaction (PCR) to amplify gene segments between the primers. A similar approach can also be applied to generate double stranded cDNA molecules after amplification of mRNA with appropriate primers and polymerases. The gene can than be isolated using a specific hybridization probe based on the amplified gene segment, which is then analyzed for appropriate expression of the protein.

The nucleotide sequence of the isolated genetic material which encodes GTBP protein can be obtained by sequencing the non-vector nucleotide sequences of these recombinant molecules. Nucleotide sequence information can be obtained by employing widely used DNA sequencing protocols, such as Maxam and Gilbert sequencing, dideoxy nucleotide sequencing according to Sanger, and the like. Examples of suitable nucleotide sequencing protocols can be found in Berger and Kimmel, *Methods in Enzymology Vol 52 Guide to Molecular Cloning Techniques*, (1987) Academic

Press. Nucleotide sequence information from several recombinant DNA isolates, including isolates from both cDNA and genomic libraries, may be combined so as to provide the entire amino acid coding sequence of GTBP, as well as the nucleotide sequences of upstream and downstream nucleotide sequences.

Nucleotide sequences obtained from sequencing GTBP protein-specific genomic library isolates can be subjected to further analysis in order to identify regions of interest in the GTBP gene. These regions of interest include additional open reading frames, promoter sequences, termination sequences, and the like. Analysis of nucleotide sequence information is preferably performed by computer. Software suitable for analyzing nucleotide sequences for regions of interest is commercially available and includes, for example, DNASIS (Pharmacia Biotech). It is also of interest to use amino acid sequence information obtained from the sequencing of purified GTBP protein when analyzing new GTBP nucleotide sequence information so as to improve the accuracy of the nucleotide sequence analysis.

Expression of GTBP

Isolated nucleotide sequences encoding GTBP protein can be used to produce purified GTBP protein or fragments thereof by either recombinant DNA methodology or by *in vitro* polypeptide synthesis techniques. By "purified" and "isolated" is meant, when referring to a polypeptide or nucleotide sequence, that the indicated molecule is present in the substantial absence of other biological macromolecules of the same type. The term "purified" as used herein preferably means at least 95% by weight, more preferably at least 99% by weight, and most preferably at least 99.8% by weight, of biological macromolecules of the same type present (but water, buffers, and other small molecules, especially molecules having a molecular weight of less than 1000, can be present).

A significant advantage of producing GBTB protein by recombinant DNA techniques rather than by isolating from natural sources of GTBP protein is that equivalent quantities of GTBP protein can be produced by using less starting material than would be required for isolating GTBP protein from a natural source. Producing GTBP protein by recombinant techniques also permits GTBP protein to be isolated in the absence of some molecules normally present in cells that naturally produce GTBP protein. It is also apparent that recombinant DNA techniques can be used to produce GTBP protein polypeptide derivatives that are not found in nature, such as the variations described above.

GTBP protein and polypeptide derivatives of GTBP protein can be expressed by recombinant techniques when a DNA sequence encoding the relevant molecule is functionally inserted into a vector. By "functionally inserted" is meant in proper reading frame and orientation, as is well understood by those skilled in the art. Typically, the GTBP protein gene will be inserted downstream from a promoter and will be followed by a stop codon, although production as a hybrid protein followed by cleavage may be used, if desired. In general, host-cell-specific sequences improving the production yield of GTBP protein and GTBP polypeptide derivatives will be used, and appropriate control sequences will be added to the expression vector, such as enhancer sequences, polyadenylation sequences, and ribosome binding sites.

Two basic types of expression are contemplated: (i) expression in mammalian cells so as to overcome a deficiency in an individual having insufficient GTBP, and (ii) expression for the purpose of providing GTBP for purpose irrelevant to the host in which expression occurs, such as production of diagnostic tests for GTBP deficiency.

Production of genetic constructs for transformation of human cells

With the goal of expression in human cells, a gene construct will be prepared and used to transform human cells. Several strategies and vectors have been developed for the expression of proteins in animal cells. For example BK-SV40 hybrid vectors have been constructed . These vectors can be maintained in cultured human cells as multicopy double-stranded DNA extrachromosomal replicons. One exemplary vector consists of the SV40 promoter controlling the expression of neomycin resistance gene (the selectable marker) and the MMTV promoter regulated by the DRE enhancer sequence which controls the expression of the cloned gene. In any case, the foreign construct will usually include transcriptional and translational initiation and termination signals, with the initiation signals 5' to the gene and termination signals 3' to the gene of interest, altough linear DNA can be delivered to a host where recombination occurs for insertion into the host genome. Expression under the control of the native promoter can thus be achieved by replacing the defective gene with the linear DNA encoding GTBP by making use of cellular processes, e.g. homologous recombination. The transcriptional initiation region which includes the RNA polymerase binding site (promoter) may be native to the host or may be derived from an alternative source, where the region is functional in the host. The transcriptional initiation regions may not only include the RNA polymerase binding site, but also regions providing for the regulation of the transcription. The 3' termination region may be derived from the same gene as the trancriptional initiation region or a different gene. For example, where the gene of interest has a trascriptional termination region functional in the host species, that region may be retained within the gene.

An expression cassette can be constructed which will include transcriptional initiation region, the GTBP protein gene under the transcriptional control of the trascription initiation region, the initiation codon, the coding sequence of the gene, with or without introns, and the translational stop codons, followed by the transcriptional termination region, which will include the terminator, and may include a polyadenylation signal sequence, and other sequences associated with transcriptional termination. The direction is 5' to 3' same as the direction of transcripition. The cassette will usually be less than about 10 kb, frequently less than about 6 kb, usually being at least about 5 kb.

When the expression product of the gene is to be located other than in the cytoplasm, the gene will usually be constructed to include particular amino acid sequences which result in translocation of the product to a particular site, which may be an organelle, such as the nucleus, or may be secreted into the external envirnoment of the cell. Various secretory leaders, membrane integrator sequences, and translocation sequences for directing the peptide expression product to a particular site are described in the literature.

One or more cassettes may be involved, where the cassettes may be employed in tandem for the expression of independent genes which may express products independently of each other or may be regulated concurrently, where the products may act independently or in conjunction, e.g. GTBP and hMSH2.

The expression cassette will normally be carried on a vector having at least one replication system. For convenience, it is common to have a replication system functional in *E. coli* such as ColE1, pSC101, pACYC184, or the like. In this manner, at each stage after each manipulation, the resulting construct may be cloned, sequenced, and the correctness of the manipulation determined. In addition, or in place of the *E. coli*

replication system, a broad host range replication system may be employed, such as the replication systems of the Pl incompatibility plasmids, e.g. RK2, RP1, RP4 and R68.

In addition to the replication system, there will frequently be at least one marker present, which may be uselful in one or more hosts, or different markers for individual hosts. That is, one marker may be employed for selection in a prokaryotic host, while another marker may be employed for selection in a eukaryotic host. Various genes which may be employed include *neo* (neomycin-kanamycin resistance), choramphenicol acetyltransferase (*cat*), b lactamase (*bla*), b galactosidase etc.

The various fragments comprising the various constructs, expression cassettes, markers, and the like may be introduced consecutively by restriction enzyme cleavage of an appropriate replication system, and insertion of the particular construct or fragment into the available size. After ligation and cloning the vector may be isolated for further manipulation. All of these techniques are amply exemplified in the literature and find particular exemplification in Maniatis *et al.*, *Molecular cloning: a laboratory manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982.

<u>Transformation of mammalian cells and gene therapy</u>

Once the vector is completed, the vector may be introduced into mammalian cells. Techniques for transforming mammalian cells include transfection, microinjection, liposome-based delivery etc.. Transfection of cultured human cells is the most commonly used method and can be achieved by standard protocols which involve either incubation of cells with DNA that has been co-precipitated with calcium phosphate or DEAE-dextran or electroporation with purified transfecting DNA. In other systems, a genetically modified virus, a liposome or a microinjection can also be used to deliver foreign DNA to human recipient cells. Once the *GTBP* gene has been introduced into the defective cell, it can

complement the genetic defect, restoring the normal phenotype. This methodology, when used to remediate genetic defects in individuals, goes under the name of gene therapy. At least two strategies for implementing somatic cell gene therapy have emerged and could be applied to correct GTBP genetic defects: *ex vivo* and *in vivo* gene therapy. Usually, the *ex vivo* gene therapy involves the following procedures:

- collect the cells from an affected individual
- correct the genetic defect by gene transfer
- select and grow the genetically corrected (remedial) cells
- infuse or transplant corrected cells back into the patient.

Vectors derived from retroviruses are often used to stably maintain and persintently express the remedial gene in the corrected cell.

In vivo gene therapy entails the direct delivery of remedial gene into the cell of a particular tissue of a prospective patient. The wild-type protein can be cloned into various benign viruses and delivered to target defective cells in an *in vivo* infection. Vectors derived from adenovirus, herpes simplex virus and certain retroviruses are excellent candidates for *in vivo* gene therapy. Methods and prospectives of gene therapy have been reviewed by Mulligan (1993), *Science* 260:926-932.

Diagnostic methods using antigens

Typically, methods for detecting analytes such as binding proteins of the invention are based on immunoassays. Immunoassays can be conducted to determine the presence or absence of GTBP in host cells. Such techniques are well known and need not be described here in detail. Examples include both heterogeneous and homogeneous immunoassay techniques. Both techniques are based on the formation of an immunological complex between the binding protein and a corresponding specific antibody. Heterogeneous assays for GTBP typically use a

specific monoclonal or polyclonal antibody bound to solid surface, e.g. in sandwich assays. Homogeneous assays, which are carried out in solution without the presence of a solid phase, can be used, for example, by determining the difference in enzyme activity brought on by binding of free antibody to an enzyme-antigen conjugate. A number of suitable assays are disclosed in U.S. Patent Nos. 3,817,837, 4,006,360, and 3,996.34545.

The solid surface reagent in the above assay prepared by known techniques for attaching protein material to solid support material, such as polymeric beads, dip sticks, or filter material. These attachment methods generally include non-specific adsorption of the protein to the support or covalent attachment of the protein, typically through a free amine group, to a chemically reactive group on the solid support, such as an activated carboxyl, hydroxyl, or aldehyde group.

In a second diagnostic configuration, known as a homogeneous assay, antibody binding to an analyte produces some change in the reaction medium which can be directly detected in the medium. Known general types of homogeneous assays proposed heretofore include (a) spin-labeled reporters, where antibody binding to the antigen is detected by a change in reported mobility (broadening of the spin splitting peaks), (b) fluorescent reporters, where binding is detected by a change in fluorescence emission, (c) enzyme reporters, where antibody binding effects enzyme/substrate interactions, and (d) liposome-bound reporters, where binding leads to liposome lysis and release of encapsulated reporter. The adaptation of these methods to the protein antigen of the present invention follows conventional methods for preparing homogeneous assay reagent.

In each of the assays described above, the assay method involves reacting the tissue extract from a test individual with an antibody and examining the sample for the presence of bound antigen. The examination may

involve attaching a labelled anti-GTBP antibody to the primary complex formed between GTBP and the immobilized antibody and measuring the amount of reporter bound to the solid support, as in the first method, or may involve observing the effect of antibody binding on a homogeneous assay reagent, as in the second method.

Production of specific binding proteins

GTBP, in its native or chemically modified form, or polypeptide derivatives thereof, or specific complexes with other polypeptides may be used for producing antibodies, either monoclonal or polyclonal, specific to GTBP or polypeptide derivatives thereof, or to GTBP complexes with other polypeptides. Antibodies specific for GTBP protein are produced by immunizing an appropriate vertebrate host, e.g., rabbit or mouse, with purified GTBP protein or polypeptide derivatives of GTBP protein, by themselves or in conjunction with a conventional adjuvant. Usually, two or more immunizations will be involved, and blood or spleen will be harvested a few days after the last injection. For polyclonal antisera, the immunoglobulins can be precipitated, isolated and purified by a variety of standard techniques, including affinity purification using GTBP protein attached to a solid surface, such as a gel or beads in an affinity column. For monoclonal antibodies, the splenocytes will normally be fused with an immortalized lymphocyte, e.g., a myeloid cell line, under selective conditions for hybridoma formation. The hybridomas can then be cloned under limiting dilution conditions and their supernatants screened for antibodies having the desired specificity. Techniques for producing antibodies are well known in the literature and are exemplified by the publication *Antibodies: A Laboratory Manual* (1988) eds. Harlow and Lane, Cold Spring Harbor Laboratories Press, and U.S. Patent Nos. 4,381,292, 4,451,570, and 4,618,577.

GTBP diagnostic application using genetic probes

The genetic material of the invention can itself be used in numerous assays as probes for genetic material present in an individual. The analyte can be a nucleotide sequence which hybridizes with a probe comprising a sequence of at least about 16 consecutive nucleotides, usually 30 to 200 nucleotides, up to substantially the full sequence of the gene as shown in SEQ ID NO: 12. The analyte can be RNA or DNA. The sample is typically a DNA or an RNA molecule extracted by the patient's tissue. In order to detect an analyte, where the analyte hybridizes to a probe, the probe may contain a detectable label. Particularly preferred for use as a probe are sequences up to about 3200 consecutive nucleotides (for example from nucleotide 1 to nucleotide 3000 of SEQ ID NO: 12 and from nucleotide 1 to nucleotide 204 of SEQ ID NO:16) since these sequences appear to be particularly specific for GTBP.

One method for amplification of target nucleic acids, for later analysis by hybridization assays, is known as the polymerase chain reaction or PCR technique. The PCR technique can be applied to detect sequences of the invention in suspected samples using oligonucleotide primers spaced apart from each other and based on the genetic sequence set forth in SEQ ID NO: 12 and SEQ ID NO:16. The primers are complementary to opposite strands of a double-stranded DNA molecule and are typically separated by from about 50 to 450 nt or more (usually not more than 2000 nt). This method entails preparing the specific oligonucleotide primers and then repeated cycles of target DNA denaturation, primer binding, and extension with a DNA polymerase to obtain DNA fragments of the expected length based on the primer spacing. Extension products generated from one primer serve as additional target sequences for the other primer. The degree of amplification of a target sequence is controlled by the number of cycles that are performed and is theoretically calculated by the simple formula 2^n where n is the number

of cycles. Given that the average efficiency per cycle ranges from about 65% to 85%, 25 cycles produce from 0.3 to 4.8 million copies of the target sequence. The PCR method is described in a number of publications, including Saiki et al., Science (1985) 230:1350-1354; Saiki et al., Nature (1986) 324:163-166; and Scharf et al., Science (1986)233:1076-1078. Also see U.S. Patent Nos. 4,683,194; 4,683,195; and 4,683,202.

The invention includes a specific diagnostic method for determination of GTBP, based on selective amplification of GTBP-protein-encoding DNA fragments. This method employs a pair of single-stranded primers derived from non-homologous regions of opposite strands of GTBP DNA duplex fragment having a sequence as described by combining the sequences SEQ ID NO: 16 and SEQ ID NO:12. These "primer fragments" represent one aspect of the invention. The method follows the process for amplifying selected nucleic acid sequences as disclosed in U.S. Patent No. 4,683,202, as discussed above.

Mutations in the *GTBP* gene can be detected by restriction enzyme analysis of the amplification product or by direct sequencing. Also, alterations in *GTBP* sequence can be revealed by Southern hybridization with probes encompassing part or the entire sequences of SEQ ID NO: 12 and SEQ ID NO:16.

Single-stranded DNA probes complementary to the wild-type GTBP-coding sequence can also be hybridized to RNA extracted from tissues or cells of human patients and used to detect mutations in the mature *GTBP* gene transcript by enzymatic digestion of heteroduplexes at the level of mismatches. These and other techniques aimed to identify variations in gene sequences from wild-type *GTBP* are extensively reported in the literature and well established in the scientific community.

Binding assays involving GTBP

a mixture of two proteins of apparent molecular weights of 100 and 160 kDa (16), which indicates that the mismatch-specific complex is composed of two proteins. The 100 kDa constituent of the complex is hMSH2 (17) while the second component is GTBP. Examples regarding the identity and function of GTBP are reported below.

Example 1

The present example shows that the GTBP protein sequence, as reported by combining the sequences SEQ ID NO:15 and SEQ ID NO: 1, contains seven subsequences which correspond to polypeptides obtained after proteolytic cleavage of the 160 kDa DNA-binding protein termed GTBP. These subsequences are indicated as SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7 and SEQ ID NO: 8. The 160 kDa protein was purified as reported in ref. 16. The fractions containing the G/T-specific mismatch binding activity were loaded onto a preparative SDS-PAGE gel and the 100 and 160 kDa bands were excised following staining with Coomassie Blue. The proteins were digested in the gel matrix either with trypsin (100 kDa protein, Promega Corporation, UK), or with *Achromombacter lyticus* endopeptidase lys-C (160 kDa protein, Wako Chemicals GmbH, Germany). The proteolytic peptides were recovered by sequential extractions and separated by tandem hplc on a Hewlett-Packard 1090M with diode array detection. Anion-exchange and octadecyl reverse phase columns were connected in series, essentially as described by H. Kawasaki and K. Suzuki, *Anal. Biochem.* **186**, 264 (1990). Fractions were collected and applied directly to an Applied Biosystems 477A pulsed-liquid automated sequencer modified as described by N.F. Totty, M.D. Waterfield and J.J. Hsuan, *Protein Sci*, 1, 1215 (1992). Microsequencing yielded seven proteolytic peptides whose sequences have been designated as SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7 and SEQ ID NO: 8.

Example 1B

The present example shows that the protein GTBP contains an amino-terminal domain corresponding to SEQ ID NO:15. This region can be determined by analysis of the coding nucleotide sequence. The amino-terminal domain is an integral part of the peptide GTBP itself, and therefore the GTBP sequence must be understood to be the sequenced combination of SEQ ID NO:15 and SEQ ID: NO:1 with a total extension of 1360 amino acids. Part a of figure 8 shows the experimental approach followed to discover the amino-terminal region of GTBP (from amino acid 1 to 68 of SEQ ID NO:15). Using the 5' RACE method(Rapid Amplification cDNA Ends, given in detail in the publication Nicolaides, N.C. et al. *Genomics*, 29: 229-234, 1995 and Nicolaides N.C. et al. *Genomics*, 30: 195-206, 1995) it is possible to determine the sequence upstream of the amino acid Ala in position 1 of SEQ ID NO:1. Initially, a pair of oligonucleotides was used that pairs with the sequence given in SEQ ID NO:12 from nucleotide 114 to 133 (primary oligonucleotide A) and from nucleotide 56 to 74 (secondary oligonucleotide B). The PCR reaction products were sequenced and it was possible to determine that the amplification product was capable of encoding the polypeptide DAAWSEAGPGPR, corresponding to amino acids 46-58 of the amino-terminal domain of GTBP as indicated in SEQ ID NO:15. Using a further two oligonucleotides, whose sequence was deduced from the initial RACE, complementary to the sequence given in SEQ ID NO:16 from nucleotide 188 to 204 (primary oligonucleotide C) and from oligonucleotide 169 to 185 (secondary oligonucleotide D) it was possible to amplify the GTBP-coding region 5' by-passing the methionine in position 1 of the amino acid sequence given in SEQ ID NO:15. The amplified clone, termed KMN, contained the entire nucleotidic sequence given in SEQ ID NO:16. RACE analysis of leucocyte cDNA is shown in lanes 2 and 5, that of placenta cDNA in lanes 3 and 6. The products of lanes 1 to 3 derive from sequenced amplifications with

oligonucleotides A and B, those in lanes 4 to 6 derive from sequenced amplifications with oligonucleotides C and D. Lanes 1 and 4 are the negative controls (absence of template). The molecular weight markers are indicated at the side.

Part b of figure 8 shows expression of the transcript encoding the protein GTBP using RT-PCR (PCR preceded by inverse transcription on RNA templates). The RT-PCR was carried out using a synthetic oligonucleotide which paired with the sequence given in SEQ ID NO:12 from nucleotide 114 to 133 in the inverse transcription reaction followed by amplification with an oligonucleotide with a sequence equal to the end 5′ of the GTBP transcript, that is 5′GGTGCTTTTAGGAGCCCCG3′.

The RNA used as a mold was taken from HeLa cells (lane 2) placenta (lane 3) leucocytes (lane 4) and cells from the colon (lane 5); these were incubated with (+ symbol on the lane) or without (- symbol on the lane) inverse transcriptase and then made to underto PCR. Wnere no cDNA was produced, as the reverse transcription reaction did not occur, it was not seen to be amplified. Lane 1 is the negative control without RNA.

<u>Example 2</u>

The present example shows that DNA regions internal to *GTBP* gene can be obtained by amplification with primers designed on the basis of the sequence of peptides deriving from proteolytic cleavage of the 160 kDa G/T-binding factor (SEQ ID NO: 2 to 8). Following the strategy of Lingner *et al.* (18) the inventors identified a unique DNA sequence encoding the central 8 amino acids of the peptide of SEQ ID NO: 6. Two degenerate primers corresponding to the N- and C-terminal amino acid sequences of the oligopeptide of SEQ ID NO: 6, i.e. the DNA sequences 5'GCGAATTCTAYGGNTTYAAYGC3' (SEQ ID NO: 9) and

5'GCGGATCCTAYTGDATNACYTC3' (SEQ ID NO: 10),

where N=any nucleotide, Y=C or T and D=A, G or T

were used for PCR amplification on poly-A^+ HeLa mRNA as described (18) except that the $MgCl_2$ concentration was 5 mM. The expected 67 bp fragment was eluted from an acrylamide gel, cloned into pBluescript SK- and sequenced (see. comments to SEQ ID NO: 9 and 10 for details). Two clones contained the correct sequence, corresponding to SEQ ID NO: 11, encoding the starting target peptide SEQ ID NO: 6..

Example 3

The present example shows that DNA regions internal to *GTBP* gene can be identified by hybridization with a DNA probe designed on the basis of the sequence of peptides obtained upon proteolytic cleavage of the 160 kDa G/T-binding factor. The DNA sequence reported as SEQ ID NO: 11 was was labeled with ^{32}P by a standard kinase reaction (with T4 PNK and [g-^{32}P]ATP as described by Maniatis *et al., Molecular cloning: a laboratory manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982) in order to generate a double-stranded DNA probe. The labelled probe of SEQ ID NO: 11 was then used in the screening of a commercial oligo dT-primed cDNA library in phage lambda (HeLa S3 Uni-ZAP XR, Stratagene). Two positive clones were selected for further analysis. Clone C1 contained an insert of 3980 bp corresponding to SEQ ID NO: 12, with a continuous open reading frame from amino acid residue 1 to 1292 encoding a polypeptide of 1292 amino acids (SEQ ID NO: 1) and a calculated molecular mass of 142 kDa; clone FLY 5 contained sequences coding from aa residue 116 to 1292 (see comments to SEQ ID NO: 1 and 12).

As all seven peptides obtained from the microsequencing of the 160 kDa protein (SEQ ID NO: 2 to 8) could be found in SEQ ID NO: 1, it can be concluded that clone C1 encodes GTBP.

Example 4

The present examples shows that GTBP protein can be used as an antigen to produce highly specific antibodies

which recognize GTBP but not hMSH2. PCR fragments corresponding to amino acid residues 27 to 158 of hMSH2 (SEQ ID NO: 13) and 750 to 928 of GTBP (SEQ ID NO: 14) were subcloned into the *E. coli* expression vector pGEX-3X (Pharmacia/LKB) and the recombinant proteins, in the form of fusion polypeptides with glutathione S-transferase, were induced and isolated as recommended by the manufacturer, except that the final concentration of IPTG was 0.25 mM and induced cultures were harvested after 6 hours at 20°C. The fusion proteins were used for immunization of New Zealand White S.P.F. female rabbits (Charles River Co.) using standard protocols. Two polyclonal antisera specifically immunoreactive to GTBP and hMSH2, respectively, were obtained and assayed as reported in *Antibodies: A Laboratory Manual* (1988) eds. Harlow and Lane, Cold Spring Harbor Laboratories Press (see Figures 2 and 5 for more details).

Example 5

The following example shows that GTBP belongs to a class of DNA-repair proteins conserved over a wide evolutionary range. Figure 3 shows the alignment of the amino acid sequences of the conserved C-terminal regions of the mismatch binding proteins GTBP (*H. sapiens*), hMSH2 (*H. sapiens*), MSH2 (*S. cerevisiae*) and MutS (*E. coli*). Identical residues are in black boxes, conserved ones in shaded boxes. Sequences reported in the alignment correspond to entries MSH2_YEAST (MSH2) and MUTS_ECOLI (MutS) in the SwissProt databank, or the coding region of GenBank entry HSU04045 (hMSH2). The alignment was carried out using the GCG Pileup option. The figure was generated using Prettyplot. The alignment reveals a high degree of conservation at the C-terminal domain among all the proteins. GTBP can thus be considered a new member of the MutS Homolog (MSH) family.

However, GTBP must be considered structurally distinct from MSH proteins, since the N-terminal domain (up to approximatively 1000 amino acids) of GTBP exhibits

remarkable divergency from MSH (human, yeast or bacterial). This is particularly evident when the homology matrixes of hMSH2 versus MSH2 (Figure 4 section d) and GTBP versus hMSH2 (Figure 4 section c) or GTBP versus MSH2 (Figure 4 section b) are compared to one another. In contrast, clear evidence is provided that GTBP is conserved over a wide evolutionary range and that structural homologs of GTBP through the whole sequence can also be found , e.g. in yeast (GenBank accession number Z47746, Figure 4 section a).

Example 6

The following example demonstrates that selective antisera recognize hMSH2 and GTBP bound to mismatched DNA in a complex. Figure 5 shows the effect of anti-hMSH2 and anti-GTBP antisera on the formation of the specific mismatch-binding complex. This gel-shift analysis was carried out as described (15), except that nuclear extracts were used (25). The antisera were added to the reaction mixtures 20 min prior to the radioactively-labelled probe. The figure is an autoradiogram of a native 6% polyacrylamide gel run in TAE buffer. Pre-incubation of the HeLa nuclear extracts with either antiserum prior to the addition of the G/T heteroduplex probe resulted in the diminuition of the specific band in a gel-shift assay, an effect not observed when the respective pre-immune sera were used. This result indicates that both proteins are present in the mismatch-specific factor. This finding also implies that extracts from cells lacking either protein are devoid of mismatch-binding activity.

Example 7

The following example shows that GTBP and hMSH2 can be expressed separately in a cell-free translation system. The inventors employed a hMSH2 cDNA clone (17) and the GTBP clones C1 and FLY5 as set forth in SEQ ID NO: 12. The C1 and FLY5 ORFs were introduced into pCite-2b. The hMSH2 ORF was inserted into pCite-1 (Novagen). *In*

vitro transcription and translation reactions were carried out as described previously (26) including a mock translation reaction in the absence of added DNA. ^{35}S-labeled translation products were analyzed on a SDS-polyacrylamide gel treated with Amplify (Amersham), dried and autoradiographed. The experiment was carried out using conditions recommended by the manufacturer. The figure is an autoradiogram of a denaturing 7.5% SDS-polyacrylamide gel. As shown in Fig. 6 section **a**, translation of hMSH2, GTBP (C1) and FLY5 mRNAs in a reticulocyte lysate system (Promega) gave rise to polypeptides of 113, 142 and 122 kDa respectively. Thus, translation of all three mRNAs gave rise to protein products of the expected size.

Example 8

The following examples shows that GTBP binds G/T mismatches when complexed to hMSH2. This was achieved by testing the two polypeptides expressed in a cell-free translation system for their ability to bind mismatch-containing substrates. Reconstitution of the mismatch-binding activity using *in vitro* translated GTBP and hMSH2 is shown in Figure 6 section b. The figure shows a gel-shift analysis showing the binding of the *in vitro*-translated proteins to the G/T heteroduplex. When GTBP and hMSH2 proteins were tested for mismatch binding activity, it was noted that expression of either protein alone has no effect on the intensity of the endogenous G/T-specific band present in the lysates at low levels. In contrast, mixing of the hMSH2 and GTBP translation products resulted in a reproducible increase in the intensity of the mismatch-specific band. This result is confirmed by using the GTBP cDNA clone FLY5, which encodes a truncated GTBP protein (see SEQ ID NO: 1 and 12). Mixing of hMSH2 and FLY5 translation products with the G/T probe gave rise to a new band with a faster electrophoretic mobility than the endogenous complex, such as would be expected of a smaller species. This

experiment provides convincing evidence that the human mismatch binding complex is composed of hMSH2 and GTBP.

Gel-shift assays were performed as described in (15). 5ml aliquots of the single *in vitro* translation reactions were tested; in the pre-mixing experiments, 2.5 ml of each of the two translation reactions were mixed and incubated for 15 min at room temperature before the addition of the probe. 5 mM AMP was included in all the DNA binding reactions so as to overcome the effect of ATP in the reticulocyte lysates, which prevents the formation of mismatch-specific protein/DNA complexes (16). The figure is an autoradiogram of a native 6% polyacrylamide gel run in TAE buffer.

Genetic alterations in mismatch repair genes such as *hMSH2, hMLH1, hPMS1* and *hPMS2* (1) are known to cause the hypermutability found in many forms of hereditary colorectal cancers (CRC). Here we report examples showing that different cell lines from CRC, which display hypermutable phenotype, contain mutated *GTBP* alleles which are expressed into non functional proteins. We also show that the spectrum of mutations found in these cell lines is different from that caused by the inactivation of *hMSH2* or of other mismatch repair genes. The following examples confirm the role of GTBP in the maintenance of human genome integrity *in vivo* and provide an explanation for the mutator phenotype observed in different CRC.

<u>Example 9</u>

The following example shows that mismatch binding activity is absent from extracts of LoVo and DLD1 cells, both derived from human CRC. LoVo cells contain a homozygous deletion in both *hMSH2* alleles (13) while neither *hMSH2* allele appears to be mutated in the cell line DLD1 (19). Extracts of LoVo and DLD1 cells fail to make mismatch-specific complexes as revealed by gel-shift assay shown in Figure 7 section a (probes were prepared as described previously (15) and experimental conditions were as in Figure 5). The figure is an autoradiogram of a

native 6% polyacrylamide gel run in TAE buffer showing the absence of specific DNA-protein complexes of expected molecular mass in both LoVo and DLD1 extracts. Based on this it appears evident that the DLD1 cell line must be devoid of GTBP. Confirmatory results were also obtained by direct screening of LoVo and DLD1 cell extracts with specific antibodies directed against GTBP and hMSH2. As expected, western blot analysis of HeLa extracts revealed the presence of equivalent amounts of hMSH2 and GTBP. In contrast, LoVo cells could be shown to lack hMSH2, and DLD1 extracts were completely devoid of full-length GTBP (Figure 7 section **b**). Interestingly, the amounts of hMSH2 in DLD1 and of GTBP in LoVo extracts were considerably lower than in the HeLa extracts. Our explanation for this finding is that hMSH2 and GTBP are unstable when not in a complex (16).

<u>Example 10</u>

The CRC-derived cell line HCT15 contains a full length hMSH2 protein but shows hypermutable phenotype (19). To determine whether HCT15 had a mutation in the GTBP coding sequence, the RNA of this cell line was reverse transcribed with random hexamers and reverse transcriptase according to standard protocols (e.g., see Powell *et al., New Engl. J. Med. 329*, 1982, 1993). The cDNA was then amplified with PCR using primers specific for the GTBP-coding sequence. The oligonucleotides used were: primer 5'-PGAGGGTTACCCCTGG-3' and 5'-ACACTGTAAGTCTGTGTACC-3' for codons 32 to 458, primers 5'-PAGTGAAGGCCTGAACAGCC-3' and 5'-AAGTCCAGTCTTTCGAGCC-3' for codons 219 to 858, and primers 5'-PGAGAGGGTTGATACTTGCC-3' and 5'-AGAAGTCAACTCAAAGCTTCC-3' for codons 692 to 1292 (where P denotes a T7 promoter sequence and a ribosome-binding site for translation initiation (26) and codon numbers are those reported in SEQ ID NO: 1 and SEQ ID NO: 12). To detect mutations in the GTBP-coding sequence, the amplification products were first transcribed and translated *in vitro* using a commercial kit (Promega).

Analysis of translation products in a PAGE-SDS gel revealed truncated GTBP polypeptides from two PCR products, corresponding to regions located at codons 32-458 (5'-end of the gene) and 692-1292 (3'-end of the gene). Sequencing of these PCR products using a commercial system (SequiTherm Polymerase, Epicentre Technologies) revealed that truncations were due to frameshift mutations. The deletion of nucleotide 664 (a C) at codon 222 changed a leucine to a termination codon and a substitution of nucleotides 3307-3312 (GATAGA) with T (see SEQ ID NO: 12) created a new termination codon several bp downstream.

Example 11

MT1 is an alkylation-resistant lymphoblastoid cell line with a biochemical deficiency similar to that of HCT15 (see Goldmacher *et al.*, *J. Biol. Chem.*, *261*, 12462, 1986; Kat *et al.* *Proc. Natl. Acad Sci USA*, *90*, 6424, 1993). To ascertain whether MT1 had a GTBP mutation, the RNA of this cell line was reverse transcribed with random hexamers and reverse transcriptase and the cDNA was then amplified with PCR using primers specific for the GTBP-coding sequence as reported above. *In vitro* transcription and translation of GTBP-coding sequence from MT1 did not reveal truncated GTBP polypeptide after electrophoretic analysis. The coding region of GTBP was therefore sequenced and two missense mutation were found in the GTBP cDNA. The first was an GAT to GTT transversion at codon 1145 of SEQ ID NO: 1, resulting in a substitution of aspartic acid with valine. The aspartic acid at codon 1145 is located in the putative DNA-binding domain of GTBP, and the identical amino acid is found at homologous positions in GTBP (*H. sapiens*), hMSH2 (*H. sapiens*), MSH2 (*S. cerevisiae*) and MutS (*E. coli*). This highly conserved amino acid residue is therefore necessary for GTBP activity and non conservative substitutions at this residue cause dramatic refuction of GTBP funcionality. The second was a GTT to ATT transition, resulting in a

substitution of isoleucine to valine at codon 1193 of SEQ ID NO: 1.

The amplification products were cloned in the vector BLUESCRIPT SK$^-$ and individual clones were sequenced using conventional protocols (Sequenase, USB). The two mutations were not found to be associated in a single clone, deriving thus from separate alleles.

Example 12

A tumor cell line, termed 543X (from the patient's designation) was derived from CRC and displays hypermutable phenotype and microsatellite instability but no mutation in mismatch repair genes so far described, including *hMSH2*, *hMLH1*, *hPMS1* and *hPMS2* (Liu *et al.*, *Nature Genetics* 9, 48, 1995). To ascertain whether 543X had a GTBP mutation, the RNA of this cell line was reverse transcribed with random hexamers and reverse transcriptase and the cDNA was then amplified with PCR using primers specific for the GTBP-coding sequence as reported above. *In vitro* transcription and translation of GTBP-coding sequence from 543X revealed truncated GTBP polypeptide after electrophoretic analysis. The sequence of the DNA region encoding GTBP was found to contain a 1 bp insertion (a T) at nucleotide 1876 of SEQ ID NO: 12, resulting in a frameshift and a downstream termination codon. The same mutation was identified also in the tumor tissue from this patient, but not in normal colon tissue. This proves that the mutation was somatic in nature and that it did not occur after the establishment of the cell culture line.

SEQUENCE LISTING

GENERAL INFORMATION

(i) APPLICANT: ISTITUTO DI RICERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S.p.A.

(ii) TITLE OF INVENTION: POLYPEPTIDE FOR REPAIRING GENETIC INFORMATION, NUCLEOTIDIC SEQUENCE WHICH CODES FOR IT AND PROCESS FOR THE PREPARATION THEREOF

(iii) NUMBER OF SEQUENCES: 16

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Societa Italiana Brevetti
(B) STREET: Piazza di Pietra, 39
(C) CITY: Rome
(D) COUNTRY: Italy
(E) POSTAL CODE: 1-00186

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk 3.5" 1.44 MBYTES
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS Rev.5.0
(D) SOFTWARE: Microsoft Word 6.0

(viii) ATTORNEY INFORMATION
(A) NAME: DI CERBO, Mario (Dr.)
(C) REFERENCE: RM/X88551/PC-DC

(ix) TELECOMMUNICATION INFORMATION
(A) TELEPHONE: 06/6785941
(B) TELEFAX: 06/6794692
(C) TELEX: 612287 ROPAT

(1) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 1292 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vi) ORIGINAL SOURCE:

(A) ORGANISM: *Homo sapiens*

(vii) IMMEDIATE SOURCE: cDNA clone pCITE2b-C1

(ix) FEATURE: SEQ ID NO: 1 shows the 1292 amino acid sequence (in three letter code) of GTBP encoded by clone C1 (see SEQ ID NO: 12). The seven oligopeptides which were identified upon proteolytic cleavage of GTBP (see SEQ ID NO: 2 to 8) are underlined. The first amino acid residue of the peptide encoded by the FLY5 *cDNA* is Asn at position 116.

(A) NAME: C1

(C) IDENTIFICATION METHOD: Experimentally

(xi) *SEQUENCE DESCRIPTION: SEQ ID NO: 1:*

Ala Lys Asn Leu Asn Gly Gly Leu Arg Arg Ser Val Ala Pro Ala Ala
1 5 10 15
Pro Thr Ser Cys Asp Phe Ser Pro Gly Asp Leu Val Trp Ala Lys Met
20 25 30
Glu Gly Tyr Pro Trp Trp Pro Cys Leu Val Tyr Asn His Pro Phe Asp
35 40 45
Gly Thr Phe Ile Arg Glu Lys Gly Lys Ser <u>Val Arg Val His Val Gln</u>
50 55 60
<u>Phe Phe Asp Asp Ser Pro</u> Thr Arg Gly Trp Val Ser Lys Arg Leu Leu
65 70 75 80
Lys Pro Tyr Thr Gly Ser Lys Ser Lys Glu Ala Gln Lys Gly Gly His
85 90 95
Phe Tyr Ser Ala Lys Pro Glu Ile Leu Arg Ala Met Gln Arg Ala Asp
100 105 110
Glu Ala Leu Asn Lys Asp Lys Ile Lys Arg Leu Glu Leu Ala Val Cys
115 120 125
Asp Glu Pro Ser Glu Pro Glu Glu Glu Glu Glu Met Glu Val Gly Thr
130 135 140
Thr Tyr Val Thr Asp Lys Ser Glu Glu Asp Asn Glu Ile Glu Ser Glu
145 150 155 160
Glu Glu Val Gln Pro Lys Thr Gln Gly Ser Arg Arg Ser Ser Arg Gln
165 170 175

Ile Lys Lys Arg Arg Val Ile Ser Asp Ser Glu Ser Asp Ile Gly Gly
180 185 190

Ser Asp Val Glu Phe Lys Pro Asp Thr Lys Glu Glu Gly Ser Ser Asp
195 200 205

Glu Ile Ser Ser Gly Val Gly Asp Ser Glu Ser Glu Gly Leu Asn Ser
210 215 220

Pro Val Lys Val Ala Arg Lys Arg Lys Arg Met Val Thr Gly Asn Gly
225 230 235 240

Ser Leu Lys Arg Lys Ser Ser Arg Lys Glu Thr Pro Ser Ala Thr Lys
245 250 255

Gln Ala Thr Ser Ile Ser Ser Glu Thr Lys Asn Thr Leu Arg Ala Phe
260 265 270

Ser Ala Pro Gln Asn Ser Glu Ser Gln Ala His Val Ser Gly Gly Gly
275 280 285

Asp Asp Ser Ser Arg Pro Thr Val Trp Tyr His Glu Thr Leu Glu Trp
290 295 300

Leu Lys Glu Glu Lys Arg Arg Asp Glu His Arg Arg Arg Pro Asp His
305 310 315 320

Pro Asp Phe Asp Ala Ser Thr Leu Tyr Val Pro Glu Asp Phe Leu Asn
325 330 335

Ser Cys Thr Pro Gly Met Arg Lys Trp Trp Gln Ile Lys Ser Gln Asn
340 345 350

Phe Asp Leu Val Ile Cys Tyr Lys Val Gly Lys Phe Tyr Glu Leu Tyr
355 360 365

His Met Asp Ala Leu Ile Gly Val Ser Glu Leu Gly Leu Val Phe Met
370 375 380

Lys Gly Asn Trp Ala His Ser Gly Phe Pro Glu Ile Ala Phe Gly Arg
385 390 395 400

Tyr Ser Asp Ser Leu Val Gln Lys Gly Tyr Lys Val Ala Arg Val Glu
405 410 415

Gln Thr Glu Thr Pro Glu Met Met Glu Ala Arg Cys Arg Lys Met Ala
420 425 430

His Ile Ser Lys Tyr Asp Arg Val Val Arg Arg Glu Ile Cys Arg Ile
435 440 445

Ile Thr Lys Gly Thr Gln Thr Tyr Ser Val Leu Glu Gly Asp Pro Ser
450 455 460

Glu Asn Tyr Ser Lys Tyr Leu Leu Ser Leu Lys Glu Lys Glu Glu Asp

465 470 475 480
Ser Ser Gly His Thr Arg Ala Tyr Gly Val Cys Phe Val Asp Thr Ser
485 490 495
Leu Gly Lys Phe Phe Ile Gly Gln Phe Ser Asp Asp Arg His Cys Ser
500 505 510
Arg Phe Arg Thr Leu Val Ala His Tyr Pro Pro Val Gln Val Leu Phe
515 520 525
Glu Lys Gly Asn Leu Ser Lys Glu Thr Lys Thr Ile Leu Lys Ser Ser
530 535 540
Leu Ser Cys Ser Leu Gln Glu Gly Leu Ile Pro Gly Ser Gln Phe Trp
545 550 555 560
Asp Ala Ser Lys Thr Leu Arg Thr Leu Leu Glu Glu Glu Tyr Phe Arg
565 570 575
Glu Lys Leu Ser Asp Gly Ile Gly Val Met Leu Pro Gln Val Leu Lys
580 585 590
Gly Met Thr Ser Glu Ser Asp Ser Ile Gly Leu Thr Pro Gly Glu Lys
595 600 605
Ser Glu Leu Ala Leu Ser Ala Leu Gly Gly Cys Val Phe Tyr Leu Lys
610 615 620
Lys Cys Leu Ile Asp Gln Glu Leu Leu Ser Met Ala Asn Phe Glu Glu
625 630 635 640
Tyr Ile Pro Leu Asp Ser Asp Thr Val Ser Thr Thr Arg Ser Gly Ala
645 650 655
Ile Phe Thr Lys Ala Tyr Gln Arg Met Val Leu Asp Ala Val Thr Leu
660 665 670
Asn Asn Leu Glu Ile Phe Leu Asn Gly Thr Asn Gly Ser Thr Glu Gly
675 680 685
Thr Leu Leu Glu Arg Val Asp Thr Cys His Thr Pro Phe Gly Lys Arg
690 695 700
Leu Leu Lys Gln Trp Leu Cys Ala Pro Leu Cys Asn His Tyr Ala Ile
705 710 715 720
Asn Asp Arg Leu Asp Ala Ile Glu Asp Leu Met Val Val Pro Asp Lys
725 730 735
Ile Ser Glu Val Val Glu Leu Leu Lys Lys Leu Pro Asp Leu Glu Arg
740 745 750
Leu Leu Ser Lys Ile His Asn Val Gly Ser Pro Leu Lys Ser Gln Asn
755 760 765

Asn Ala Ala Arg Leu Ala Asn Leu Pro Glu Glu Val Ile Gln Lys Gly
1235 1240 1245

His Arg Lys Ala Arg Glu Phe Glu Lys Met Asn Gln Ser Leu Arg Leu
1250 1255 1260

Phe Arg Glu Val Cys Leu Ala Ser Glu Arg Ser Thr Val Asp Ala Glu
1265 1270 1275 1280

Ala Val His Lys Leu Leu Thr Leu Ile Lys Glu Leu
1285 1290

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 10 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vi) ORIGINAL SOURCE:

(A) ORGANISM: Homo sapiens

(ix) FEATURE: SEQ ID NO: 2 to 8 show seven oligopeptides derived from proteolytic cleavage of GTBP extracted from HeLa cells and purified as described in ref. 16 . The peptide corresponding to SEQ ID NO: 6 (18 amino acids) was selected to design two degenerate primers corresponding to the N- and C-terminal sequences of the peptide, as given in detail in SEQ ID NO: 9 and 10.

(A) NAME: FR44

(C) IDENTIFICATION METHOD: Experimentally

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

Val Arg Val His Val Gln Phe Phe Asp Asp
1 5 10

(3) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 18 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR48
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

Lys Leu Pro Asp Leu Glu Arg Leu Leu Ser Lys Ile His Asn Val XXX
1 5 10 15
Ser Lys

(4) INFORMATION FOR SEQ ID NO: 4:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 13 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR49b
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

Leu Ser Arg Gly Iso Gly Val Met Leu Pro Gln Val Leu
1 5 10

(5) INFORMATION FOR SEQ ID NO: 5:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 14 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

The presence of an altered GTBP protein can be detected by the use of binding assays based on the specific recognition of G/T mismatches by GTBP. A synthetic double-stranded 34-mer oligonucleotide containing G/T mispair is prepared and labelled substantially as reported (15). Cell extracts can be prepared as reported in current literature (e.g. ref 25 and refs. therein). The cell extract (1-10 micrograms of nuclear proteins) can be incubated with the heteroduplex oligonucleotide at room temperature for 30 minutes to allow GTBP binding to the G/T mismatch. The mixture can then be loaded on a gel prepared as reported in Figure 6. Alterations in GTBP mass or affinity for the substrate can be evidenced by an altered electrophoretic mobility.

Deposits

Strains of E. coli TOP10 - transformed using the plasmids pBluescript SK^-/C1 and pCite-2b/C1 coding respectively for the protein GTBP from the amino acid 1 to the amino acid 1292 of SEQ ID NO:1 and using the plasmid pBluescript SK^-/FLYS coding for a GTBP protein from the amino acid 116 to the amino acid 1292 of SEQ ID NO:1 - have been deposited on 19/5/1995 with the National Collections of Industrial and Marine Bacteria Ltd. (NCIMB), Aberdeen,Scotland, UK, with accession numbers NCIMB 40742, NCIMB 40471 and NCIMB 40740 respectively. Moreover, a strain of E.coli TOP10 - transformed using the plasmid pBluescript SK^-/GTBP coding for the whole amino acid sequence of GTBP from the amino acid 1 to the amino acid 1360 (SEQ ID NO: 15 and SEQ ID NO:1) - has been deposited on 28/5/96 with the above depositary institution with accession number NCIMB 40805.

Examples

As mentioned above, the inventors identified a mismatch-binding factor in HeLa cells (15), GTBP, which was shown to bind preferentially to heteroduplexes containing G/T mispairs. Purification of this DNA binding activity by G/T mismatch affinity chromatography yielded

(ii) MOLECULE TYPE: protein
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR49c
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

```
Thr Leu Arg Thr Leu Leu Glu Glu Glu Tyr Phe Arg Glu Lys
1               5                   10
```

(6) INFORMATION FOR SEQ ID NO: 6:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 18 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: HeLa cell extract
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR52
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

```
Ser Tyr Gly Phe Asn Ala Ala Arg Leu Ala Asn Leu Pro Glu Glu Val
1               5                   10                  15
Ile Gln
```

(7) INFORMATION FOR SEQ ID NO: 7:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 13 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein

(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR59
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

```
Asn Pro Glu Gly Arg Phe Pro Asp Leu Thr Val Glu Leu
1               5                   10
```

(8) INFORMATION FOR SEQ ID NO: 8:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 11 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(ix) FEATURE: see SEQ ID NO: 2
(A) NAME: FR69
(C) IDENTIFICATION METHOD: Experimentally
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

```
Ile Ile Asp Phe Leu Ser Ala Leu Glu Gly Phe
1               5                   10
```

(9) INFORMATION FOR SEQ ID NO: 9
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: synthetic DNA
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vii) IMMEDIATE SOURCE: oligonucleotide synthesizer

(ix) FEATURE: SEQ ID NO:9 shows the sequence of the degenerate single-stranded DNA primer deduced from the N-terminal of oligopeptide shown in SEQ ID NO: 6. Together with SEQ ID NO: 10, the two primers were used to amplify poly-A^+ RNA extracted from HeLa cells. The expected 67 base pairs (bp) fragment was cloned in pBluescript SK^- (Stratagene) and sequenced with a commercial T7-polymerase based kit (Pharmacia). The 54 bp sequence of the resulting fragment, obtained after subtraction of the engineered cloning sites, is shown as SEQ ID NO: 11.

(A) NAME: oligo 5' sense

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9

GCGAATTCTA YGGNTTYAAY GC 22

(10) INFORMATION FOR SEQ ID NO: 10

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 22 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: Yes

(vii) IMMEDIATE SOURCE: oligonucleotide synthesizer

(ix) FEATURE: SEQ ID NO:10 shows the sequence of the degenerate single-stranded DNA primer deduced from the C-terminal of oligopeptide shown in SEQ ID NO: 6. Together with SEQ ID NO: 9, the two primers were used to amplify poly-A^+ RNA extracted from HeLa cells. The expected 67 base pairs (bp) fragment was cloned in pBluescript SK^- (Stratagene) and sequenced with a commercial T7-polymerase based kit (Pharmacia).The 54 bp sequence of the resulting fragment, obtained

after subtraction of the engineered cloning sites, is shown as SEQ ID NO: 11.

(A) NAME: oligo 3' antisense

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10

GCGGATCCTC YTGDATNACY TC 22

(11) INFORMATION FOR SEQ ID NO: 11

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 54 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: Yes

(vii) IMMEDIATE SOURCE: PCR product

(ix) FEATURE: SEQ ID NO: 11 shows the double-stranded DNA sequence encoding the oligopeptide reported in SEQ ID NO: 6, as deduced by sequencing of cloned amplification product. This fragment was derived from PCR amplification of HeLa cDNA, using the degenerate primers described in SEQ ID NO: 9 and 10. The DNA sequence was end-labelled with ^{32}P by a standard kinase reaction (with T4 polynucleotide kinase PNK and [g-^{32}P]ATP as described by Maniatis *et al.*, *Molecular cloning: a laboratory manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982) in order to generate a double-stranded DNA probe. The labelled probe was used in the screening of a commercial oligo dT-primed cDNA library in phage lambda (HeLa S3 UNI-ZAP XR, Stratagene). Screening of the HeLa S3 UNI-ZAP XR library in phage lambda made it possible the identification of two clones hybridizing with the DNA probe. These clones were designated C1 and FLY5.

(A) NAME:

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11

AGCTATGGCT TTAATGCAGC AAGGCTTGCT AATCTCCCAG AGGAAGTTAT TCAA

54

(12) INFORMATION FOR SEQ ID NO: 12

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 3980 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vii) IMMEDIATE SOURCE: cDNA clone C1

(ix) FEATURE: SEQ ID NO: 12 shows the 3980 bp cDNA sequence of clone C1. The cDNA insert of clone FLY5 spanned from nucleotide 346 to 3980 of the C1 sequence as reported in SEQ ID NO: 12.

(A) NAME: C1

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12

```
GCGAAGAACC TCAACGGAGG GCTGCGGAGA TCGGTAGCGC CTGCTGCCCC CACCAGTTGT    60
GACTTCTCAC CAGGAGATTT GGTTTGGGCC AAGATGGAGG GTTACCCCTG GTGGCCTTGT   120
CTGGTTTACA ACCACCCCTT TGATGGAACA TTCATCCGCG AGAAAGGGAA ATCAGTCCGT   180
GTTCATGTAC AGTTTTTTGA TGACAGCCCA ACAAGGGGCT GGGTTAGCAA AAGGCTTTTA   240
AAGCCATATA CAGGTTCAAA ATCAAAGGAA GCCCAGAAGG GAGGTCATTT TTACAGTGCA   300
AAGCCTGAAA TACTGAGAGC AATGCAACGT GCAGATGAAG CCTTAAATAA AGACAAGATT   360
AAGAGGCTTG AATTGGCAGT TTGTGATGAG CCCTCAGAGC CAGAAGAGGA AGAAGAGATG   420
GAGGTAGGCA CAACTTACGT AACAGATAAG AGTGAAGAAG ATAATGAAAT TGAGAGTGAA   480
GAGGAAGTAC AGCCTAAGAC ACAAGGATCT AGGCGAAGTA GCCGCCAAAT AAAAAAACGA   540
AGGGTCATAT CAGATTCTGA GAGTGACATT GGTGGCTCTG ATGTGGAATT TAAGCCAGAC   600
ACTAAGGAGG AAGGAAGCAG TGATGAAATA AGCAGTGGAG TGGGGGATAG TGAGAGTGAA   660
GGCCTGAACA GCCCTGTCAA AGTTGCTCGA AAGCGGAAGA GAATGGTGAC TGGAAATGGC   720
TCTCTTAAAA GGAAAAGCTC TAGGAAGGAA ACGCCCTCAG CCACCAAACA AGCAACTAGC   780
ATTTCATCAG AAACCAAGAA TACTTTGAGA GCTTTCTCTG CCCCTCAAAA TTCTGAATCC   840
CAAGCCCACG TTAGTGGAGG TGGTGATGAC AGTAGTCGCC CTACTGTTTG GTATCATGAA   900
ACTTTAGAAT GGCTTAAGGA GGAAAAGAGA AGAGATGAGC ACAGGAGGAG GCCTGATCAC   960
```

CCCGATTTTG ATGCATCTAC ACTCTATGTG CCTGAGGATT TCCTCAATTC TTGTACTCCT 1020
GGGATGAGGA AGTGGTGGCA GATTAAGTCT CAGAACTTTG ATCTTGTCAT CTGTTACAAG 1080
GTGGGGAAAT TTTATGAGCT GTACCACATG GATGCTCTTA TTGGAGTCAG TGAACTGGGG 1140
CTGGTATTCA TGAAAGGCAA CTGGGCCCAT TCTGGCTTTC CTGAAATTGC ATTTGGCCGT 1200
TATTCAGATT CCCTGGTGCA GAAGGGCTAT AAAGTAGCAC GAGTGGAACA GACTGAGACT 1260
CCAGAAATGA TGGAGGCACG ATGTAGAAAG ATGGCACATA TATCCAAGTA TGATAGAGTG 1320
GTGAGGAGGG AGATCTGTAG GATCATTACC AAGGGTACAC AGACTTACAG TGTGCTGGAA 1380
GGTGATCCCT CTGAGAACTA CAGTAAGTAT CTTCTTAGCC TCAAAGAAAA AGAGGAAGAT 1440
TCTTCTGGCC ATACTCGTGC ATATGGTGTG TGCTTTGTTG ATACTTCACT GGGAAAGTTT 1500
TTCATAGGTC AGTTTTCAGA TGATCGCCAT TGTTCGAGAT TTAGGACTCT AGTGGCACAC 1560
TATCCCCCAG TACAAGTTTT ATTTGAAAAA GGAAATCTCT CAAAGGAAAC TAAAACAATT 1620
CTAAAGAGTT CATTGTCCTG TTCTCTTCAG GAAGGTCTGA TACCCGGCTC CCAGTTTTGG 1680
GATGCATCCA AAACTTTGAG AACTCTCCTT GAGGAAGAAT ATTTTAGGGA AAAGCTAAGT 1740
GATGGCATTG GGGTGATGTT ACCCCAGGTG CTTAAAGGTA TGACTTCAGA GTCTGATTCC 1800
ATTGGGTTGA CACCAGGAGA GAAAAGTGAA TTGGCCCTCT CTGCTCTAGG TGGTTGTGTC 1860
TTCTACCTCA AAAAATGCCT TATTGATCAG GAGCTTTTAT CAATGGCTAA TTTTGAAGAA 1920
TATATTCCCT TGGATTCTGA CACAGTCAGC ACTACAAGAT CTGGTGCTAT CTTCACCAAA 1980
GCCTATCAAC GAATGGTGCT AGATGCAGTG ACATTAAACA ACTTGGAGAT TTTTCTGAAT 2040
GGAACAAATG GTTCTACTGA AGGAACCCTA CTAGAGAGGG TTGATACTTG CCATACTCCT 2100
TTTGGTAAGC GGCTCCTAAA GCAATGGCTT TGTGCCCCAC TCTGTAACCA TTATGCTATT 2160
AATGATCGTC TAGATGCCAT AGAAGACCTC ATGGTTGTGC CTGACAAAAT CTCCGAAGTT 2220
GTAGAGCTTC TAAAGAAGCT TCCAGATCTT GAGAGGCTAC TCAGTAAAAT TCATAATGTT 2280
GGGTCTCCCC TGAAGAGTCA GAACCACCCA GACAGCAGGG CTATAATGTA TGAAGAAACT 2340
ACATACAGCA AGAAGAAGAT TATTGATTTT CTTTCTGCTC TGGAAGGATT CAAAGTAATG 2400
TGTAAAATTA TAGGGATCAT GGAAGAAGTT GCTGATGGTT TTAAGTCTAA AATCCTTAAG 2460
CAGGTCATCT CTCTGCAGAC AAAAAATCCT GAAGGTCGTT TTCCTGATTT GACTGTAGAA 2520
TTGAACCGAT GGGATACAGC CTTTGACCAT GAAAAGGCTC GAAAGACTGG ACTTATTACT 2580
CCCAAAGCAG GCTTTGACTC TGATTATGAC CAAGCTCTTG CTGACATAAG AGAAAATGAA 2640
CAGAGCCTCC TGGAATACCT AGAGAAACAG CGCAACAGAA TTGGCTGTAG GACCATAGTC 2700
TATTGGGGGA TTGGTAGGAA CCGTTACCAG CTGGAAATTC CTGAGAATTT CACCACTCGC 2760
AATTTGCCAG AAGAATACGA GTTGAAATCT ACCAAGAAGG GCTGTAAACG ATACTGGACC 2820
AAAACTATTG AAAAGAAGTT GGCTAATCTC ATAAATGCTG AAGAACGGAG GGATGTATCA 2880
TTGAAGGACT GCATGCGGCG ACTGTTCTAT AACTTTGATA AAAATTACAA GGACTGGCAG 2940
TCTGCTGTAG AGTGTATCGC AGTGTTGGAT GTTTTACTGT GCCTGGCTAA CTATAGTCGA 3000
GGGGGTGATG GTCCTATGTG TCGCCCAGTA ATTCTGTTGC CGGAAGATAC CCCCCCCTTC 3060
TTAGAGCTTA AAGGATCACG CCATCCTTGC ATTACGAAGA CTTTTTTTGG AGATGATTTT 3120
ATTCCTAATG ACATTCTAAT AGGCTGTGAG GAAGAGGAGC AGGAAAATGG CAAAGCCTAT 3180

TGTGTGCTTG TTACTGGACC AAATATGGGG GGCAAGTCTA CGCTTATGAG ACAGGCTGGC 3240
TTATTAGCTG TAATGGCCCA GATGGGTTGT TACGTCCCTG CTGAAGTGTG CAGGCTCACA 3300
CCAATTGATA GAGTGTTTAC TAGACTTGGT GCCTCAGACA GAATAATGTC AGGTGAAAGT 3360
ACATTTTTTG TTGAATTAAG TGAAACTGCC AGCATACTCA TGCATGCAAC AGCACATTCT 3420
CTGGTGCTTG TGGATGAATT AGGAAGAGGT ACTGCAACAT TTGATGGGAC GGCAATAGCA 3480
AATGCAGTTG TTAAAGAACT TGCTGAGACT ATAAAATGTC GTACATTATT TTCAACTCAC 3540
TACCATTCAT TAGTAGAAGA TTATTCTCAA AATGTTGCTG TGCGCCTAGG ACATATGGCA 3600
TGCATGGTAG AAAATGAATG TGAAGACCCC AGCCAGGAGA CTATTACGTT CCTCTATAAA 3660
TTCATTAAGG GAGCTTGTCC TAAAAGCTAT GGCTTTAATG CAGCAAGGCT TGCTAATCTC 3720
CCAGAGGAAG TTATTCAAAA GGGACATAGA AAAGCAAGAG AATTTGAGAA GATGAATCAG 3780
TCACTACGAT TATTTCGGGA AGTTTGCCTG GCTAGTGAAA GGTCAACTGT AGATGCTGAA 3840
GCTGTCCATA AATTGCTGAC TTTGATTAAG GAATTATAGA CTGACTACAT TGGAAGCTTT 3900
GAGTTGACTT CTGACCAAAG GTGGTAAATT CAGACAACAT TATGATCTAA TAAACTTTAT 3960
TTTTTAAAAA TGAAAAAAAA
3980

(13) INFORMATION FOR SEQ ID NO: 13

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 394 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vii) IMMEDIATE SOURCE: Homo sapiens

(ix) FEATURE: SEQ ID NO: 13 shows the double-stranded DNA sequence used to express an internal domain of hMSH2 (corresponding to amino acid residues 27 to 158) in the expression vector pGEX-3x (see also legend to Figure 2).

(A) NAME: GST/hMSH2

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13

GGAGAAGCCG ACCACCACAG TGCGCCTTTT CGACCGGGGC GACTTCTATA CGGCGCACGG 60
CGAGGACGCG CTGCTGGCCG CCCGGGAGGT GTTCAAGACC CAGGGGGTGA TCAAGTACAT 120
GGGGCCGGCA GGAGCAAAGA ATCTGCAGAG TGTTGTGCTT AGTAAAATGA ATTTTGAATC 180
TTTTGTAAAA GATCTTCTTC TGGTTCGTCA GTATAGAGTT GAAGTTTATA AGAATAGAGC 240

TGGAAATAAG GCATCCAAGG AGAATGATTG GTATTTGGCA TATAAGGCTT CTCCTGGCAA 300
TCTCTCTCAG TTTGAAGACA TTCTCTTTGG TAACAATGAT ATGTCAGCTT CCATTGGTGT 360
TGTGGGTGTT AAAATGTCCG CAGTTGATGG CCAG 394

(14) INFORMATION FOR SEQ ID NO: 14

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 534 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vii) IMMEDIATE SOURCE:

(ix) FEATURE: SEQ ID NO: 14 shows the double-stranded DNA sequence used to express an internal domain of GTBP (corresponding to amino acid residues 750 to 928) in the expression vector pGEX-3x (see also legend to Figure 2).

(A) NAME: GST/GTBP

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14

CTTGAGAGGC TACTCAGTAA AATTCATAAT GTTGGGTCTC CCCTGAAAGT CAGAACCACC 60
CAGACAGCAG GGCTATAATG TATGAAGAAA CTACATACAG CAAGAAGAAG ATTATTGATT 120
TTCTTTCTGC TCTGGAAGGA TTCAAAGTAA TGTGTAAAAT TATAGGGATC ATGGAAGAAG 180
TTGCTGATGG TTTTAAGTCT AAAATCCTTA AGCAGGTCAT CTCTCTGCAG ACAAAAAATC 240
CTGAAGGTCG TTTTCCTGAT TTGACTGTAG AATTGAACCG ATGGGATACA GCCTTTGACC 300
ATGAAAAGGC TCGAAAGACT GGACTTATTA CTCCCAAAGC AGGCTTTGAC TCTGATTATG 360
ACCAAGCTCT TGCTGACATA AGAGAAAATG AACAGAGCCT CCTGGAATAC CTAGAGAAAC 420
AGCGCAACAG AATTGGCTGT AGGACCATAG TCTATGGATT GGTAGGAACC GTTACGCAGC 480
TGGAAATTCC TGAGAATTTC ACCACTCGCA ATTTGCCAGA AGAATACGAG TTGA 534

(15) INFORMATION FOR SEQ ID NO: 15

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 68 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens
(vii) IMMEDIATE SOURCE: cDNA of clone KMN
(ix) FEATURE: SEQ ID NO: 15 shows the amino-terminal sequence of 68 amino acids of GTBP encoded by the clone TASNR2A1 (see SEQ ID NO:16 for the corresponding nucleotide encoding sequence). The amino acid sequence SEQ ID NO:15 (corresponding to residues 1-68) must be placed in front of the amino acid in position 1 of the sequence given in SEQ ID NO:1 (corresponding to 1292 residues) to obtain the complete GTBP sequence of 1360 amino acids.
(A) NAME: KMN
(C) IDENTIFICATION METHOD: experimental
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15

```
Met Ser Arg Gln Ser Thr Leu Tyr Ser Phe Phe Pro Lys Ser Pro Ala
1               5                   10                  15
Lys Ser Asp Ala Met Lys Ala Ser Ala Arg Ala Ser Arg Glu Gly Gly
            20                  25                  30
Arg Ala Ala Ala Ala Pro Glu Ala Ser Pro Ser Pro Gly Gly Asp Ala
        35                  40                  45
Ala Tyr Ser Glu Ala Gly Pro Gly Pro Arg Pro Leu Ala Arg Ser Ala
    50                  55                  60
Ser Pro Pro Lys
65
```

(16 INFORMATION FOR SEQ ID NO: 16
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 204 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: synthetic DNA
(iii) HYPOTHETICAL: No
(iv) ANTISENSE: No

(vii) IMMEDIATE SOURCE: cDNA of clone KMN

(ix) FEATURE: SEQ ID NO: 16 shows the double-stranded DNA sequence obtained using the RACE method (Rapid Anmplification cDNA Ends) used to establish the 5'-terminal sequence of GTBP cDNA encoding the amino-terminal region of the protein GTBP as indicated in SEQ ID NO:15. The nucleotidic sequence SEQ ID NO:15 (corresponding to 204 residues) must be positioned in front of the nucleotide in position 1 of the sequence given in SEQ ID NO:12 (corresponding to 3980 residues) in order to obtain the complete GTBP-encoding sequence of 4080 nucleotides.

(A) NAME: KMN

(C) IDENTIFICATION METHOD: Polyacrylamide gel

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16

```
ATGTCGCGAC AGAGCACCCT GTACAGCTTC TTCCCCAACT CTCCGGCGCT GAGTGATGCC    60
AACAAGGCCT CGGCCAGGGC CTCACGCGAA GGCGGCCGTG CCGCCGCTGC CCCCGAGGCC   120
TCTCCTTCCC CAGGCGGGAA TGCGGCCTGG AGCGAGGCTG GGCCTGGGCC CAGGCCCTTG   180
GCGCGATCCG CGTCACCGCC CAAG    204
```

CLAIMS

1. An isolated polypeptide, wherein said polypeptide comprises: (1) a first sequence corresponding to GTBP as set forth by combining the amino acid sequences set forth in SEQ ID NO: 15 and SEQ ID NO:1; a second sequence wherein said second sequence is a subsequence of said first sequences and is at least 4 amino acids; (3) a third sequence in which at least one amino acid is replaced by a different amino acID

2. The polypeptide of Claim 1 complexed to a second polypeptide.

3. The polypeptide complex of Claim 2, wherein said second polypeptide is hMSH2.

4. An isolated polypeptide according to claim 1, comprising the amino acid sequences from amino acid 1 to 68 of SEQ ID NO:15 and from amino acid 1 to 1292 of SEQ ID NO: 1, or in any case sequences within the combination of SEQ ID NO: 15 and SEQ ID NO:1, for example SEQ ID NO: 2 to SEQ ID NO:8).

5. An isolated DNA or RNA molecule, wherein said molecule comprises:

(1) a first sequence encoding GTBP as set forth by combining SEQ ID NO:16 and SEQ ID NO: 12;

(2) a second sequence, wherein said second sequence is a subsequence of said first sequence and is at least 10 nucleotides in length;

(3) a third sequence in which at least one nucleotide of said first or second sequence is replaced by a different nucleotide; or

(4) a fourth sequence complementary to any of said first second, or third sequences;

with the provisos that (1) if said molecule is an RNA molecule, U replaces T in said sequence of said molecule, (2) said third sequence is at least 95% identical to said first or second sequence, and (3) said second sequence is not present in hMSH2 cDNA.

6. The molecule of Claim 5, wherein said molecule comprises said first sequence.

7. The molecule of Claim 5, wherein said molecule comprises said second sequence.

8. The molecule of Claim 5, wherein said molecule comprises said third sequence.

9. The molecule of Claim 5, wherein said molecule comprises a cDNA sequence.

10. The molecule of Claim 5, wherein said molecule consists essentially of DNA encoding GTBP.

11. The molecule of Claim 5, wherein the RNA or DNA encoding GTBP is naturally occuring.

12. An expression vector containing the molecule of Claim 5.

13. A cell transformed with the molecule of Claim 5.

14. The cell of Claim 13, wherein said molecule is DNA and said DNA is arranged in operative association with an expression control sequence capable of directing replication and expression of said DNA.

15. The cell according to Claim 13, wherein said cell is a eukaryotic or prokaryotic cell including animal, fungal or bacterial cell.

16. A process for producing GTBP protein comprising culturing a cell of Claim 13 in a suitable culture medium and isolating said GTBP protein from said cell.

17. A polypeptide made according to the process of Claim 16.

18. A method for identifying agents which inhibit or enhance GTBP activity as detectable by *in vitro* multi- or dimeriation assays, DNA-binding assays and mismatch repair assays.

19. A method of identifying GTBP-modulating agents, comprising:

(1) performing a heterodimerization that includes a GTBP polypeptide, hMSH2 and an agent, and (2)

detecting whether the agent modulates hetero-dimerization.

20. The method of Claim 19, wherein the heteodimerization assay comprises an *in vitro* binding reaction.

21. A preparation of specific antibodies immunoreactive with GTBP and not substantially immunoreactive with other proteins unrelated to GTBP.

22. A method of purification of GTBP or GTBP-complexing molecules involving the use of specific antibodies of Claim 21.

23. A method of purification of GTBP or GTBP-complexing molecules based on specific interaction between GTBP and nucleic acid recognition sequences.

24. A method of detecting the presence of a genetic defect that has the potential of causing tumorigenesis in human, which comprises:

identifying a mutation of a *GTBP* gene of said human, wherein said mutation results in a *GTBP* gene sequence different from wild-type human GTBP-coding DNA sequence as set forth by combining SEQ ID NO:16 and SEQ ID NO: 12.

25. A method of detecting the presence of a genetic defect that causes cancer in a human, which comprises:

identifying a mutation of a *GTBP* gene of said human, wherein said mutation provides a *GTBP* gene sequence different from human *GTBP* DNA sequence as set forth by combining SEQ ID NO:16 and SEQ ID NO: 12, that changes the sequence of a protein product of said *GTBP* gene, or that causes the GTBP product to be truncated or that results in said *GTBP* gene not being transcribed or translated.

26. A method of diagnosing or prognosing a neoplastic tissue of a human comprising:

identifying the presence of a mutation of a *GTBP* gene or its expression product in said tissue of said human patient, wherein said mutation provides a *GTBP*

gene sequence different from human *GTBP* DNA sequence as set forth by combining SEQ ID NO:12 and SEQ ID NO: 16, said alteration indicating neoplasia of the tissue.

27. The methods of Claims 24-26, wherein said mutations result in a change in the sequence of a protein product of said *GTBP* gene.

28. The methods of Claims 24-26, wherein said mutations result in said *GTBP* gene not being transcribed or translated.

29. The methods of Claims 24-26, wherein said mutations create stop codons in said *GTBP* gene.

30. The methods of Claims 24-26, wherein said methods comprise Polymerase Chain Reaction (PCR) amplification of at least a segment of said *GTBP* gene.

31. The methods of Claims 24-26, whereas said methods comprise identifying a change in a restriction site as a result of said mutation.

32. The methods of Claim 24-26, wherein said methods comprise restriction fragment length polymorphism analysis, allele-specific oligonucleotide hybridization or nucleotide sequencing.

33. The methods of Claims 24-26, wherein said methods classify said human as homozygous for said *GTBP* gene or for said mutated *GTBP* gene or heterozygous for said *GTBP* gene and said mutated *GTBP* gene.

34. The methods of Claims 24-26 wherein the expression products are mRNA molecules.

35. The methods of Claims 24-26 wherein the loss of wild-type GTBP coding sequence is detected by Nothern hybridization of mRNA molecules extracted from cells or tissues.

36. The methods of Claims 24-26 wherein the loss of wild-type *GTBP* is detected by Southern hybridization of a GTBP DNA probe to genomic DNA of said human patient.

37. The methods of Claims 24-26 wherein the loss of wild-type *GTBP* gene is detected by identifying a mismatch between nucleic acids including (1) mRNA molecules of

said human patient and (2) a nucleic acid complementary to human wild-type GTBP coding sequence, when molecules 1 and 2 are hybridized with each other and form a duplex.

38. The methods of Claims 24-26 wherein the loss of wild-type gene is detected by gene cloning and sequencing of cloned DNA.

39. The methods of Claims 24-26 wherein the loss of wild-type *GTBP* gene is detected by screening for point mutations and deletion or insertion mutations.

40. The method of Claims 24-26 wherein the expression products are protein molecules.

41. The methods of Claims 24-26 wherein the loss of wild-type GTBP is detected by immunoblotting, e.g. Western blotting.

42. The methods of Claims 24-26 wherein the alteration of wild type GTBP is detected by immunoenzymology and immunocytochemistry.

43. The method of Claims 24-26 wherein the alteration of wild-type *GTBP* is detected by binding interactions between said GTBP protein and a second cellular protein.

44. The method of Claim 43 wherein the second cellular protein is hMSH2.

45. A method for generating transgenic animals carrying mutant *GTBP* alleles.

46. A pharmaceutical composition useful in the treatment of GTBP-dependent diseases comprising a therapeutically effective amount of GTBP in a pharmaceutically acceptable vehicle.

47. A method for supplying wild-type *GTBP* gene function to a cell which has altered GTBP, said gene function being lost by virtue of a mutation in a *GTBP* gene comprising:

introducing full-length or part of *GTBP* gene in a cell which has lost such gene function such that said full-length or part of *GTBP* gene are expressed in the cell and encode full-length or part of the GTBP protein

which is capable of complementing the genetic defect at the basis of neoplastic disease.

48. A method for supplying wild-type *GTBP* gene function to a cell which has altered GTBP, said gene function being lost by virtue of a mutation in a *GTBP* gene comprising introducing into a cell a molecule which mimics the effect of GTBP alone or complexed with other molecules.

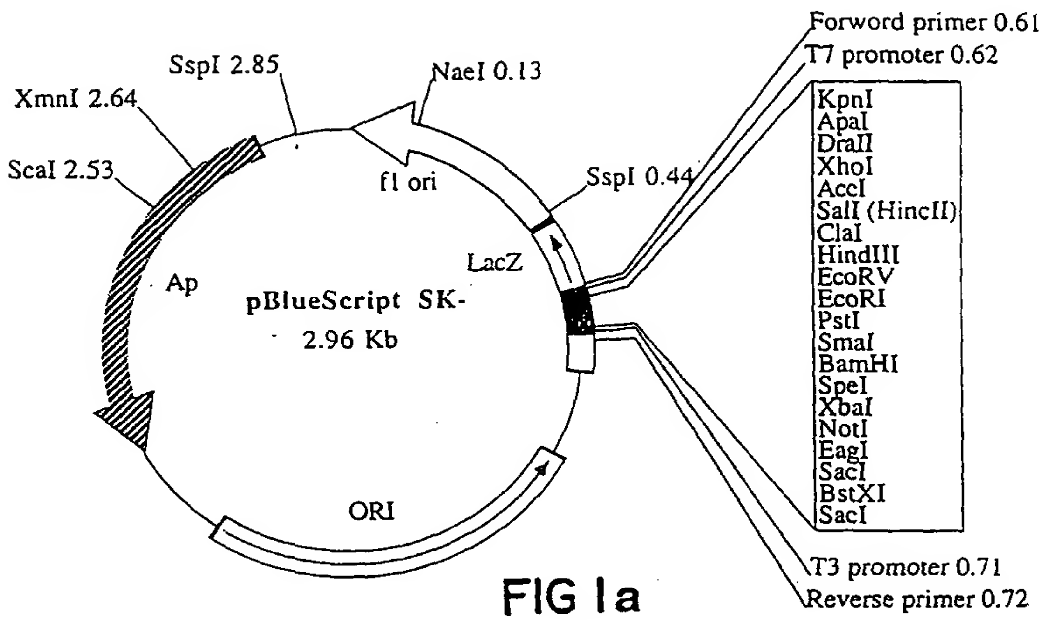

Forword primer 0.61
T7 promoter 0.62
KpnI
ApaI
DraII
XhoI
AccI
SalI (HincII)
ClaI
HindIII
EcoRV
EcoRI
PstI
SmaI
BamHI
SpeI
XbaI
NotI
EagI
SacI
BstXI
SacI
T3 promoter 0.71
Reverse primer 0.72
SspI 2.85
NaeI 0.13
XmnI 2.64
ScaI 2.53
f1 ori
SspI 0.44
LacZ
Ap
pBlueScript SK-
2.96 Kb
ORI

FIG 1a

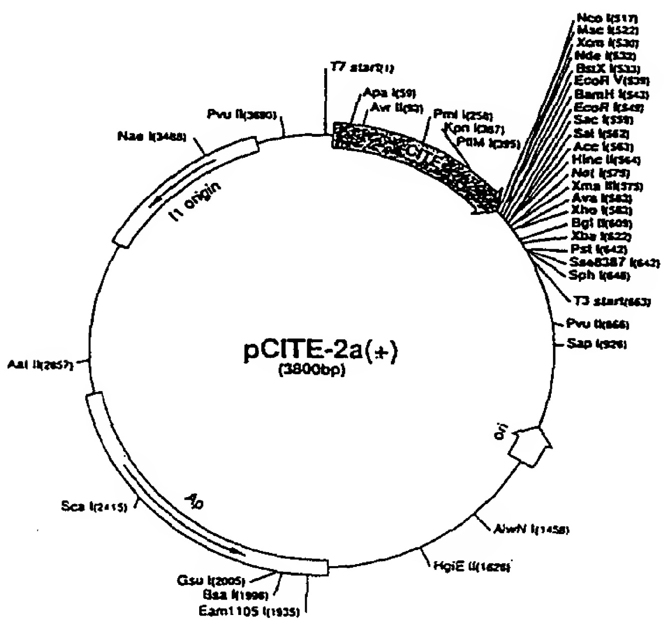

T7 start(1)
Apa I(59)
Kpn I(367)
Nae I(3488)
f1 origin
Nco I(517)
EcoR V
BamH I
EcoR I
Sac I
Sal I
Acc I
Pst I(642)
Sse8387 I(642)
Sph I(648)
T3 start(663)
Sap I(926)
pCITE-2a(+)
(3800bp)
ori
Aat II(2657)
Ap
Sca I(2415)
AlwN I(1458)
HgiE II(1628)
Gsu I(2005)
Bsa I(1996)
Eam1105 I(1935)

FIG 1b

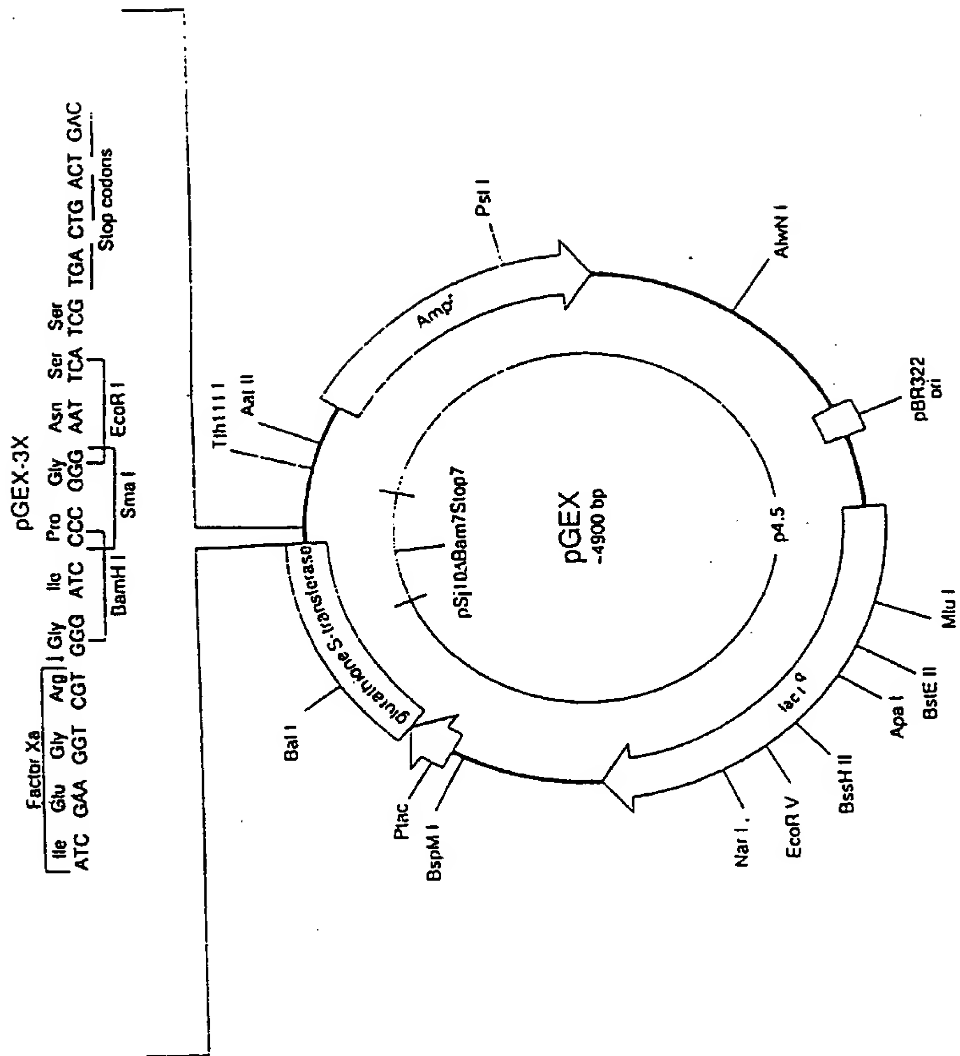

pGEX-3X
Factor Xa
Ile Glu Gly Arg Gly Ile Pro Gly Asn Ser Ser
ATC GAA GGT CGT GGG ATC CCC GGG AAT TCA TCG TGA CTG ACT GAC
BamH I
Sma I
EcoR I
Stop codons
Pst I
AlwN I
pBR322 ori
Mlu I
BstE II
Apa I
BssH II
EcoR V
Nar I
BspM I
Ptac
Bal I
Tth111 I
Aat II
Amp^r
pGEX
~4900 bp
p4.5
lac I^q
glutathione S-transferase
pSj10ΔBam7Stop7

FIG 2

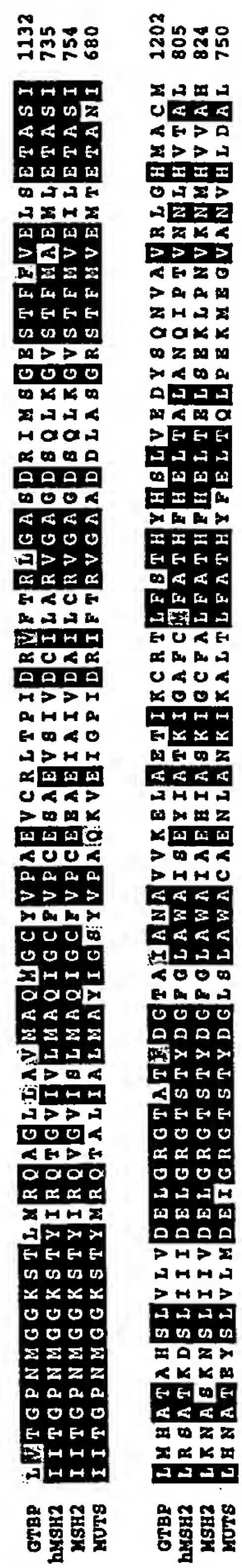


FIG 3

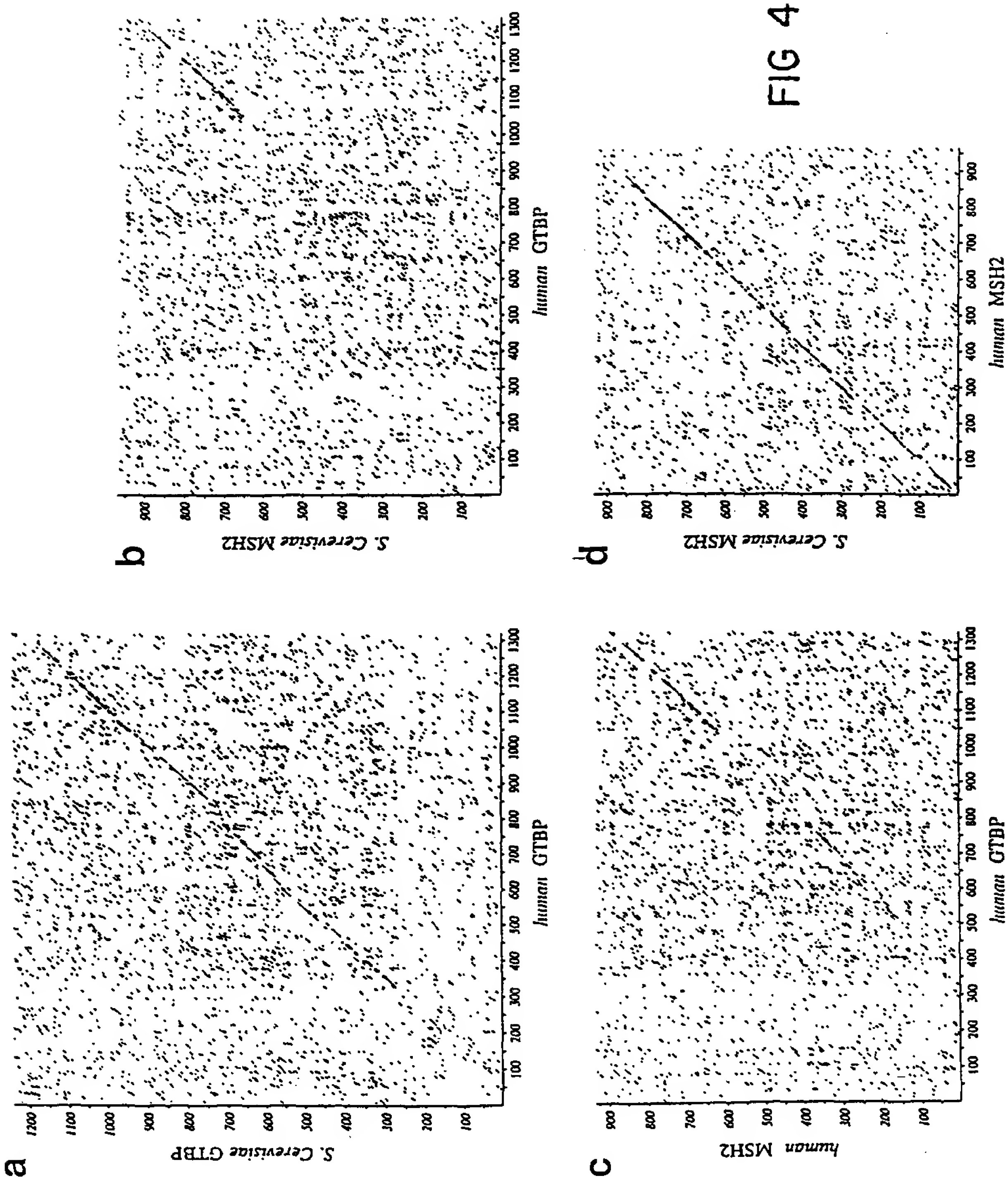

a
S. Cerevisiae GTBP
human GTBP
b
S. Cerevisiae MSH2
human GTBP
c
human MSH2
human GTBP
d
S. Cerevisiae MSH2
human MSH2

FIG 4

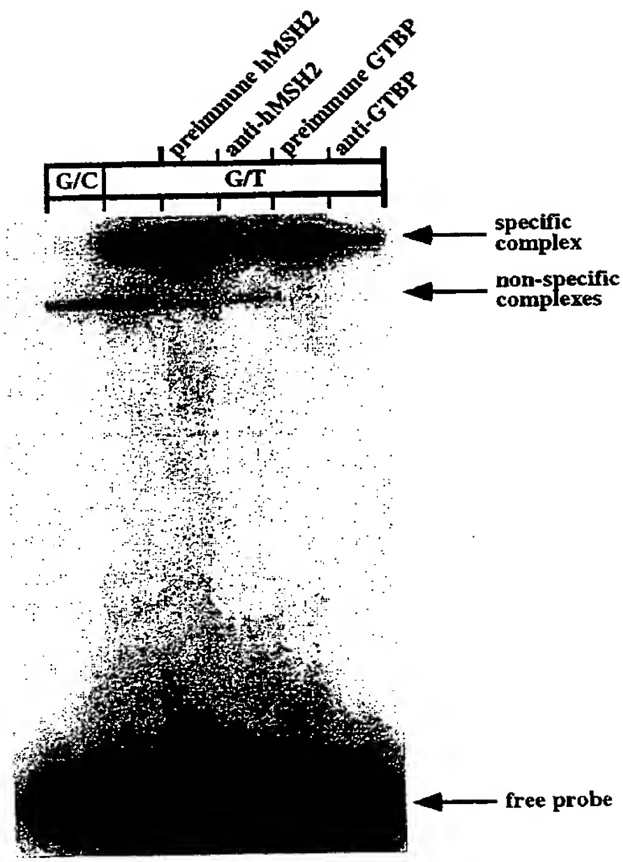

preimmune hMSH2
anti-hMSH2
preimmune GTBP
anti-GTBP
G/C
G/T
specific complex
non-specific complexes
free probe

FIG 5

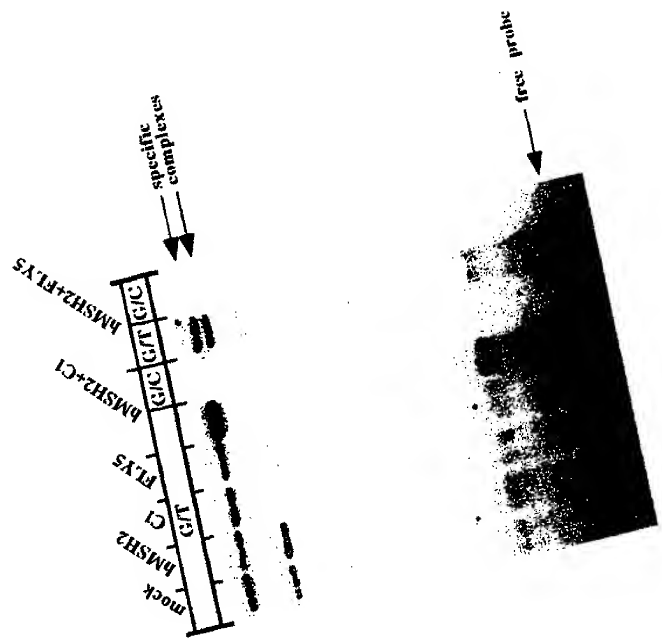

mock
hMSH2
C1
FLY5
hMSH2+C1
hMSH2+FLY5
G/T
G/C
G/T
G/C
specific complexes
free probe

FIG 6b

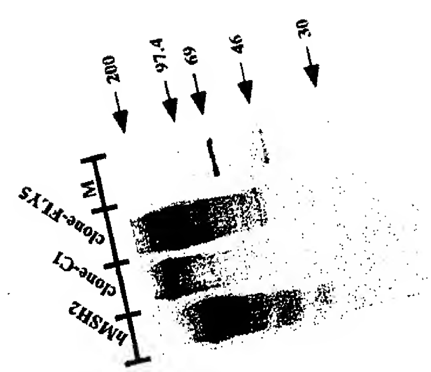

hMSH2
clone-C1
clone-FLY5
M
200
97.4
69
46
30

FIG 6a

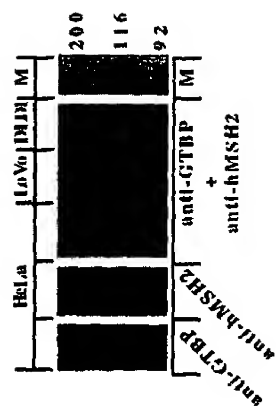

HeLa
LoVo
DLD1
M
200
116
92
anti-GTBP
anti-hMSH2
anti-GTBP
+
anti-hMSH2
M

FIG 7b

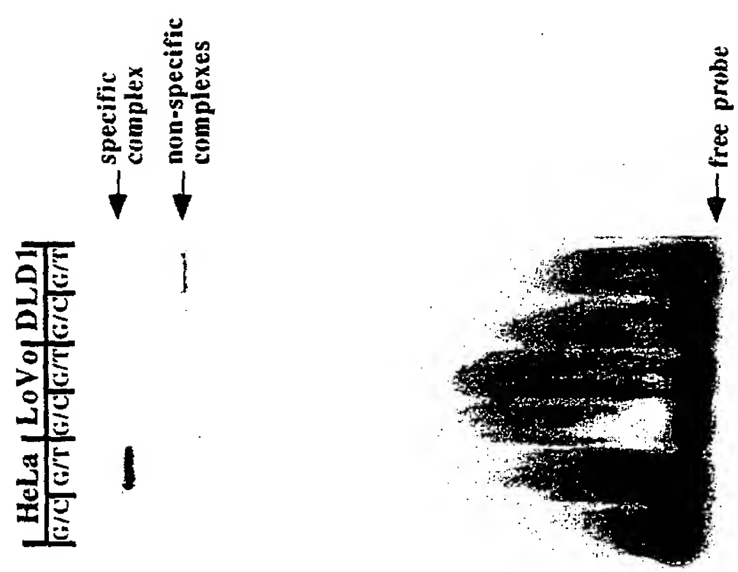

HeLa
LoVo
DLD1
G/C
G/T
G/C
G/T
G/C
G/T
specific complex
non-specific complexes
free probe

FIG 7a

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inte   nal Application No
PCT/US 96/12900

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
WO-A-9305430	18-03-93	US-A-	5218652	08-06-93
		CA-A-	2111091	18-03-93
		EP-A-	0694180	31-01-96
		JP-T-	6510140	10-11-94
US-A-4665529	12-05-87	DE-A-	3715600	03-12-87
		FR-A-	2598863	20-11-87
		GB-A,B	2190784	25-11-87
		JP-A-	62274788	28-11-87
		US-A-	4723257	02-02-88
		US-E-	RE34192	09-03-93
US-A-5127068	30-06-92	EP-A-	0486175	20-05-92
		JP-A-	5093828	16-04-93
		US-A-	5436990	25-07-95